

We have been warned

The world is celebrating the news that the SARS outbreak now seems to be contained. But the epidemic has revealed gaps in our defences against emerging viral diseases and the ever-looming threat of a flu pandemic.

It's official: on 5 July the World Health Organization (WHO) announced that the global outbreak of severe acute respiratory syndrome (SARS) had finally been contained. The chain of local transmission of SARS in Taiwan, the last region to bring the disease under control, seems to have been broken.

But the WHO rightly muted the euphoria with a warning that SARS has not actually gone away. About 200 people remain sick; others may be infected but not showing symptoms; and the SARS virus could hop over from its natural animal host at any time to haunt us once again. Global vigilance, the WHO argued, must be kept up. Indeed, the reprieve we are now celebrating may be no more than a breathing space. SARS could turn out to be seasonal, like influenza, and come back with the next Northern Hemisphere winter. If so, the intense level of scientific research into the disease must also be maintained, to develop therapies and vaccines.

The announcement that SARS has been contained provides an appropriate landmark against which to ask some searching questions about the epidemic (see page 121). There are several stages to bringing a newly emerging disease under control. First, a cluster of unusual symptoms must be recognized and public-health authorities alerted. These officials must respond by isolating patients, tracing their contacts and controlling their movement. Meanwhile, the pathogen involved must be identified. In the longer term, therapeutic strategies must be developed and the source of the infection — usually animals that harbour the virus — must be pinned down.

Rapid response

A certain amount of backslapping is warranted. Although a more infectious disease would probably have been impossible to contain, our success in controlling SARS was a consequence of rigorous international activity in isolating infected people. And when it came to characterizing the pathogen, the scientific response was exemplary. The credit goes to the network of virologists put together by the WHO, which quickly identified the SARS virus. The unsung hero was Klaus Stöhr, the WHO official who built the SARS team and nurtured an atmosphere in which normal scientific competition could be suspended. Stöhr, who leads the WHO's influenza project, built his team around the WHO's existing flu network. Even as the SARS crisis was at its peak, this network quietly reinforced its merit by identifying the virus responsible for a virulent type of avian flu that had broken out in poultry farms in the Netherlands and was jumping over into humans (see *Nature* **422**, 247; 2003).

Having been taught the merits of a network culture, we should encourage its continuation, so that when the next deadly virus starts to circulate, scientific networks of various types — for basic and clinical research, and the development of drugs and vaccines — can be activated immediately.

On other fronts, however, the SARS outbreak has revealed serious shortcomings in our ability to respond to emerging diseases. Even now, there is an urgent need for more studies to find the natural animal reservoir from which the SARS virus emerged. Without this knowledge, it will remain impossible to assess the likelihood of recurrent outbreaks.

Perhaps the biggest chink in our armour is the initial stage of clinical surveillance for unusual clusters of disease. Had China been more attentive to this, and launched a vigorous investigation into the outbreak of atypical pneumonia that emerged in its southern Guangdong province in November last year, SARS might have been nipped in the bud before it established its initial human foothold. Unfortunately, disease surveillance has never had a high priority in health spending. Although it is reasonably sophisticated in the United States, Canada and some other developed countries, it is very patchy elsewhere.

Under surveillance

In the wake of SARS, some experts are suggesting that surveillance for emerging diseases should extend to sampling and characterization of the entire panoply of viruses circulating in people and animals (see *Nature* **423**, 471; 2003). One model is the existing WHO influenza network, which samples the flu viruses in general circulation, to monitor for dangerous variants and to help plan vaccine production. Extending this approach to other viruses would be extremely expensive, however. Given the difficulty of getting politicians to spend large sums of money on averting unknown, future threats, the immediate priority should instead be on improving basic clinical surveillance. This already has a strong foundation in the form of the WHO's Global Outbreak Alert and Response Network. Established in 2000, this network involves 145 countries and aims to monitor and investigate rumours about unusual disease symptoms. But its effectiveness is limited by the fact that it merely coordinates existing, poorly resourced national surveillance activities.

We also have to face the grisly fact that even the most powerful surveillance and the most rapid scientific responses will not be enough to rapidly contain a disease that is as deadly as SARS but as infectious as influenza. But with better all-round preparedness — from better surveillance to a guaranteed capacity for vaccine production — the fight could be made more even, reducing the scale of the tragedy that would result.

The next disease that fits this description will probably not be entirely novel, but could be a new and deadly strain of influenza. Given this, the desultory state of our preparedness for a flu pandemic is shameful. SARS was a genuinely new disease, so drugs and vaccines couldn't immediately be brought into play. In contrast, effective drugs that target the flu virus do exist, and methods for the rapid production of vaccines are already established. Yet no government has stockpiled drugs, and only Canada has organized capacity for producing flu vaccines for all of its citizens. With the WHO's flu network doing such a sterling job in monitoring for potential pandemic strains, this failure to prepare for converting the knowledge it provides into widespread immunization is shockingly negligent.

The 1918 flu pandemic killed tens of millions, and we know for certain that a strain with a similar combination of virulence and infectiousness will at some point cross over again from animals to people. The SARS outbreak should be seen as a timely warning to health officials who have yet to make adequate preparations for this eventuality. This time, we had a lucky escape. ■

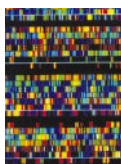




No refuge
US farmers breaking
the rules on
transgenic crops
p116



Nature survey
Scientific misconduct
in the German media
fuels public distrust
p117



Chips down
Affymetrix thinks
again on microarray
upgrade
p119



Deep loss
Japanese give up
the search for
missing submarine
p120

Science shunted to the sidelines as EU sets out its space agenda

Declan Butler, Paris

Plans for a sweeping revamp of Europe's space sector are in danger of neglecting space science, researchers say.

Space scientists are worried that a much-anticipated white paper, or policy document, on space strategy, to be issued in October by the European Commission, will emphasize military and commercial applications in space at the expense of science.

They also fear that the European Space Agency (ESA), which most scientists view as a force for good, could be marginalized when the European Union (EU) starts to take a lead in space policy.

Europe's activities in space — damaged by a flagging commercial satellite market and crashes of its flagship Ariane launcher — are gaining fresh political impetus this year. On top of the white paper, the draft constitution for the EU, drawn up in June by a convention chaired by former French president Valéry Giscard d'Estaing, gives responsibility for space to the EU.

Together, these events bring the EU's political clout to bear for the first time in a sector that has previously been managed by national governments and by ESA, which was created as an intergovernmental agency quite separate from the EU.

The white paper will identify an autonomous space programme as a key element of Europe's political and economic independence. This is broadly expected to open the door to more public spending on space, greater integration of civilian and military space programmes, and the expanded use of space technologies in agriculture, telecommunications and other policy areas.

Space scientists welcome all this. But at a consultative meeting in Paris on 23–24 June, many expressed concern that early drafts of the white paper relegate space science to the sidelines. Roger-Maurice Bonnet, executive director of the International Space Science Institute in Bern, Switzerland, and former head of space science at ESA, told the meeting that the white paper should be amended to include a broad vision for space science in Europe.

Jean-Louis Fellous, a member of the



Up in smoke: the new Ariane rocket failed just three minutes after its launch last October.

European Space Science Committee of the European Science Foundation, said that science should be explicitly included in the white paper's call for a doubling of Europe's total investment in space by 2010. According to the European Commission, Europe spent some €6 billion (US\$7 billion) on space last year, with 90% of this on civilian programmes, whereas the United States spent \$36 billion, split equally between civil and military spending.

Europe spends less than \$1 billion per year on space science — a sixth of the United States' investment, according to Paris-based analysts Euroconsult. "New investment in launchers and industry must not be to the detriment of science," said Fellous. European Commission officials at the conference promised scientists that their message had been heard and that the white paper would be modified accordingly.

"The white paper is a unique opportunity in the history of European space," says Bonnet. Alluding to Europe's inability until now to close the gap in funding with the United States, he argues that: "There is a profound crisis; the existing system doesn't work; so we

need to change the system."

Bonnet says that shifting responsibility for space to the European level will ultimately benefit science, as present funding is too fragmented. Integration of space within military, environmental and other EU policy sectors will result in a funding boost for space research, he predicts.

But in changing the system, scientists are also concerned that the European Commission may eventually usurp ESA in running the science. They cherish ESA's bottom-up approach, says Fellous, exemplified by its track record of backing investigator-driven proposals based on scientific excellence, and are wary of the commission's legendary bureaucracy and top-down programming of research areas.

And in a thinly veiled warning against any such takeover bid, Antonio Rodotà, director general of ESA, stressed the agency's past achievements. "We must not lose sight of what we have already accomplished, and not undo the solid foundation that has made Europe the major player it is today. Change is a must, but continuity also has its place."

Pest resistance feared as farmers flout rules

Tom Clarke, London

Nearly one-fifth of farmers in the US mid-west are ignoring federal rules about how much transgenic maize (corn) they can plant, according to government figures. Experts fear that this non-compliance could encourage insects to develop resistance to the insecticide produced by the crop.

Some transgenic maize contains a gene from the bacterium *Bacillus thuringiensis* (*Bt*), which allows the corn to produce a natural insecticide. Under rules laid down by the Environmental Protection Agency, farmers who plant *Bt* maize must devote 20% of their acreage to non-*Bt* varieties. These 'refuge' areas should prevent pests from developing resistance to the insecticide, as resistant insects will breed with susceptible insects living in the refuge and dilute the trait.

But a study released last month by the Center for Science in the Public Interest (CSPI), a Washington-based pressure group, reveals that this rule is being ignored. The report describes data from the US Department of Agriculture showing that last year 19% of all *Bt* maize-growing farms in Iowa, Minnesota and Nebraska failed to plant the necessary refuges. No refuge at all was planted on 13% of the farms.

Most farms that broke the rule were small, planting less than 80 hectares (200 acres) of *Bt* maize. Farms of this size are not monitored by the Agricultural Biotechnology Stewardship



Bt maize is resistant to the European corn borer.

Technical Committee, the industry body that monitors growing practices.

"It's just a matter of time before resistance develops," warns the report's author Gregory Jaffe, director of biotechnology projects at the CSPI. *Bt* maize will be rendered useless, adds Jaffe, if pests such as the European corn borer, the chief target of transgenic maize, develop resistance.

But transgenic-crop firms disagree. "This won't jeopardize the long-term effectiveness of the technology," says Eric Sachs, director of scientific affairs at biotechnology company Monsanto, based in St Louis, Missouri. Those ignoring the rules grow just 5% of *Bt* maize planted in the three states, he says.

Only 25% of maize grown in the midwest contains *Bt*, and the likelihood that non-compliance from a fraction of these farms will lead to resistance is small, agrees insect ecologist Bruce Tabashnik at the University of Arizona in Tucson. "It's unlikely that there is sufficient pressure on the insects," he says. In seven years of academic field surveys, insects resistant to *Bt* maize have not been documented in the United States, Tabashnik adds.

But this could change. New *Bt* maize, developed by Monsanto and designed to control a root pest, the maize rootworm, is due to be sown next year, and is expected to be adopted far more widely than existing *Bt* varieties. "Although non-compliance may not be an immediate hazard, it could lead to a serious problem in the future," says Tabashnik.

The transgenic-crop developers maintain that small farms won't pose a major problem — but they are concerned. "It's not clear that we have a problem of biological significance," says Val Giddings, vice-president for food and agriculture with the US Biotechnology Industry Organization, "but we do have a problem of regulatory significance." ■

Europe finds transgenic food hard to swallow

Jim Giles, London

Transgenic food is unlikely to appear on European tables soon, analysts say — despite a vote by the European Parliament that opens the way for European nations to declare currently unlicensed varieties of the foods safe to eat.

On 2 July, the parliament passed a measure that would require food with 0.9% or more of genetically modified content to be labelled as such. Transgenic ingredients will have to be traceable to the farm on which they were grown.

The measure, which is expected to be approved by the Council of Ministers in the next few weeks and pass into law, should open the way for the approval of new varieties of transgenic crops.

About 20 varieties, including insecticide- and herbicide-resistant maize (corn), are currently awaiting approval by the European Union (EU). The process has been blocked by countries such as France and Germany, creating a *de facto* moratorium. They and others are expected to drop their opposition

now that the new rules are agreed.

But even if the approvals process opens up, trade analysts say that sales are likely to be slow. Consumer surveys suggest that European shoppers won't buy food that is labelled as containing transgenic material, and several leading supermarket chains have refused to stock it and are expected to continue to do so.

Trevor Young, an agricultural economist at the University of Manchester, has analysed how much money consumers would need to save for them to buy transgenic crops. He says that no thorough trial has been conducted, but that preliminary studies by him and colleagues suggest that a reduction of around 40% would be needed. The cost savings associated with transgenic agriculture are unlikely to produce such big discounts.

US farmers' groups have also complained that the cost of implementing the EU's traceability rules will price their transgenic crops out of the market.

The European Parliament's vote was partly intended to stave off a pending US



claim to the World Trade Organization (WTO) that EU restrictions on the sale of transgenic food are illegal (see *Nature* 423, 369; 2003). But US officials gave no indication that the case, which some observers interpret as a warning to nations outside Europe, including larger grain importers in Asia, will be dropped. ■

Survey reveals mixed feelings over scientific misconduct

CORBIS



Close scrutiny: two-thirds of scientists have experienced misconduct, and a quarter say it is increasing.

Alison Abbott and Phillip Graf, Munich

Senior scientists in Germany believe that the increased reporting of research misconduct in the media there is damaging public confidence in science, a survey conducted by *Nature* has found.

Many acknowledge continuing problems in maintaining good scientific practice and handling allegations of misconduct, particularly in protecting 'whistleblowers' who report misconduct. But two in five said it is "unrealistic" to demand strict adherence to good scientific practice in all circumstances.

The survey, the first to test researchers' opinions after a stream of well-publicized scientific fraud cases in Germany, was conducted among top researchers who serve as reviewers for the DFG, the country's main granting agency. Of 211 researchers contacted in all scientific disciplines, 77 responded.

The DFG has taken a lead in setting standards for scientific practice and ensuring adoption of these and of investigative procedures. Some of its officials support a 'zero tolerance' policy in which all misconduct would be investigated and punished.

But 40% of respondents say deviations could be accepted in some circumstances. For example, says one, "honorary authorship is justified in some individual cases".

Some two-thirds said that they had had personal experience of misconduct, either directly or indirectly. Most felt that it is a major problem in clinical research (80%) and the life sciences (59%), but only 4% felt that this is the case in physics and chemistry.

A quarter said they thought that the overall incidence of misconduct is increasing. Eighty-five per cent said that the DFG system is helpful in exposing fraud and protecting

the rights of accused scientists. But one-third said that whistleblowers are inadequately protected. Nearly half said that newspaper reports create "an atmosphere of distrust", although a similar proportion believe that journalists play a useful role.

In German universities, researchers can confide their suspicions either to a local ombudsman at their institution, or to one of the DFG's three independent ombudsmen. These can start enquiries, either by the university itself or by a DFG-appointed panel.

DFG ombudsman Hans-Heinrich Trute, a law professor at the University of Hamburg, says he doesn't know why some young scientists apparently do not trust his office, taking their concerns to newspapers instead.

Holger Wormer, a journalist at Munich's leading newspaper the *Süddeutsche Zeitung*, who has covered several cases, says whistleblowers come to him because they fear that top scientists will protect each other. "It helps that we provide a bit of pressure to break down these old-boy networks," he says.

Trute agrees that more could be done to protect whistleblowers. In Germany, they have to prove that they are victims of malevolent reprisals, he says: the burden of proof should be shifted to those in power to prove they are not, as in the United States. But he sympathizes with scientists who voice doubts about 'zero tolerance' policies. "Sometimes intuition requires a bit of licence," he says.

Peter Hofschneider, emeritus director at the Max Planck Institute for Biochemistry at Martinsried near Munich, who has acted as an informal confidant in several cases, agrees. Young scientists "should not necessarily have their whole research career destroyed because of one misdemeanour", he says. ■

Europe split over move to loosen stem-cell regulations

Alison Abbott, Munich

The European Union (EU) is set to change its rules on embryonic-stem-cell research.

Under proposals due to be announced this week, researchers would be able to use EU funds to carry out studies of stem cells derived from embryos left over from *in vitro* fertilization. EU policy is currently unclear, deterring researchers from applying for funding.

The proposal, expected to be made when the European Commission's directorate heads meet on 9 July, will be controversial. Attitudes to embryo research differ widely among EU member states, and negotiators expect a period of intense horse-trading to take place before a vote of EU member nations in September.

Under EU rules, votes will be distributed according to population, and 29% of them need to be cast against the proposal for it to be rejected. Germany and Italy each hold 11.5% of the votes, putting the two countries in key positions.

Both might be expected to block the proposal. Germany only allows work on human embryonic stem cells from lines banked before January 2002, when its policy was set. Italy has no specific law to regulate this kind of research, but there is strong public opposition to it.

Observers say, however, that German negotiators are working behind the scenes to ensure that the proposal is endorsed. This will stop Germany's rules being imposed on other countries. Britain, for example, lets researchers create new stem-cell lines from embryos. With German public opposition to ES-cell research running high, its delegates will not openly back the proposal. But they may try to get Italy to abstain, or consider doing so themselves.

In contentious EU votes, the country that holds the rotating presidency — currently Italy — traditionally remains neutral. Together with Germany's unwillingness to scupper the proposal, this may persuade Italy to abstain.

"As Italy holds the presidency, it would be irresponsible of it to vote against," says Cinzia Caporale, a bioethicist at the University of Siena and vice-president of the committee of bioethics at the United Nations Educational, Scientific and Cultural Organization. "It would make no difference to the results — Germany would abstain if Italy insists on voting against." ■

Drive for patent-free innovation gathers pace

Declan Butler, Paris

A group of top scientists and economists are asking the World Intellectual Property Organization (WIPO) in Geneva to promote 'open' models of innovation that don't rely on patents.

The group believes that innovation based on freely available knowledge can be effective not just in areas where it has established a foothold — such as genome sequence data — but also in sectors where patent protection is entirely dominant, such as drug development (see *Nature* 424, 10–11; 2003).

In a 7 July letter to Kamil Idris, director-general of the WIPO, 59 scientists and economists call attention to the "explosion of open and collaborative projects to create public goods" in recent years, including the Human Genome Project, the open-source software movement, and Internet standards. Such projects show that "one can achieve a high level of innovation in some areas of the modern economy without intellectual property protection", says the letter, arguing that



Kamil Idris is being asked to assess the merits of an open approach to intellectual property.

"excessive, unbalanced or poorly designed intellectual property protections may be counterproductive". It calls on the WIPO to hold a major conference on these models during 2004.

The signatories include Joseph Stiglitz of Columbia University in New York, who received the 2001 Nobel prize for economics; John Sulston of the Wellcome Trust Sanger Institute near Cambridge, UK, winner of the 2002 Nobel prize for medicine; James Orbinski, former president of Médecins Sans Frontières; and Richard Stallman, a computer scientist regarded by many as the 'father' of the open-source software movement.

Francis Gurry, an assistant director-general at the WIPO, said that the organization welcomed the idea. "The use of open and collaborative development models for research and innovation is a very important and interesting development," he said in a statement. "The director-general looks forward with enthusiasm to taking up the invitation to organize a conference to explore the scope and application of these models."

Advocates of open-source innovation want the WIPO and other public agencies to rethink how innovation works, says James Love, director of the Washington-based Consumer Project on Technology and a signatory to the letter. Open research for drug development is one of the initiative's main targets, he says. Some of the authors are also pursuing the idea of an international treaty to encourage governments to fund drug research and put the results directly into the public domain.

Love argues that research results should ultimately become a freely available commodity, with drug companies competing to market generics of any drugs developed. The current system, in which drug research and development is carried out by drug companies that keep patent rights for up to 20 years, is grossly inefficient and results in excessive prices so that those who need the drugs most cannot afford them, argues Love.

Yet to be fleshed out are details of how such a model would work, and how competitive forces could be maintained within it. But in May, the general assembly of the World Health Organization instructed agency officials to draft terms of reference during 2004 for a new evaluation of intellectual property, innovation and public health. Consideration of open-science models is expected to be part of this exercise.

"The success of the Internet and of open-source software has driven home just how far open and collaborative projects can go," says Hal Varian, an economist at the University of California, Berkeley, who has also signed the 7 July letter.

Another signatory, Paul David, an economist at Stanford University, argues that systems such as free and open-source software are not at odds with intellectual property rights protection, but rather a choice by creators and society as to the benefits they want to obtain.

Malaysia puts biovalley under wraps

David Cyranoski, Kuala Lumpur

The Malaysian government is quietly launching a major project to harness the country's abundant natural biodiversity to create a viable biotechnology industry.

Over the next three years, it plans to invest 600 million ringgits (US\$160 million) to build three research institutes, dedicated to molecular biology, plant biotechnology and drug development. Based at an 80-hectare campus in Dengkil, 45 kilometres south of Kuala Lumpur, the institutes are scheduled to open in 2006.

But in contrast to similar initiatives elsewhere, Bio Valley Malaysia is being set up in an atmosphere that borders on secrecy. Although Malaysian biologists welcome the initiative, most won't discuss it on the record.

Even government officials don't want to talk about the project, which was announced

by Prime Minister Mahathir bin Mohamad in May. "We want to get the project off the ground rather than talk about it," says an official at the science ministry.

A university researcher, who didn't want to be named, says that Mahathir wants to keep the project low profile, after a comparable information-technology initiative in 1997 failed to meet its advertised goals. The government is also sensitive to charges that it is allowing private investors to exploit Malaysia's rich biodiversity.

This biodiversity distinguishes the project from other biotechnology parks, says the ministry spokesperson. Three companies have already agreed to locate themselves in the park and the government is negotiating with another 20.

Researchers are already tapping into the country's biodiversity. A collaboration between several Malaysian institutes and the Massachusetts Institute of Technology, for example, is looking at the plant tongkat ali — used in Malaysia to treat impotency.

The government hopes that the biovalley will enable firms to develop drugs locally by bioprospecting in Malaysia. But some analysts doubt whether Malaysian investors are ready to back long-term, research-intensive ventures, such as biotech companies. And researchers say that bureaucratic restrictions on their mobility, and on the efficient filing of patents, will make it difficult for a biotechnology sector to take root.



Malaysia hopes to build a biotech industry on the back of its rainforest biodiversity.

DNA-chip firm backs down over upgrade

John Whitfield, London

Biotechnology company Affymetrix is retreating from a plan that would have made it tougher for biologists to use open-source computer software to analyse results from its ubiquitous microarray chips.

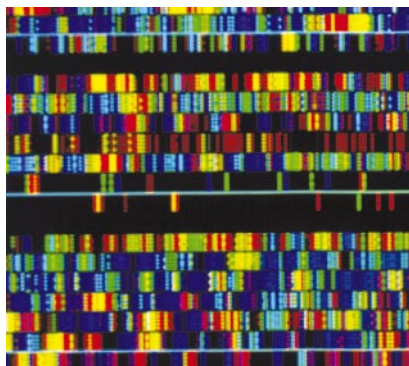
The firm, based in Santa Clara, California, had planned to upgrade its microarrays, which are used by thousands of molecular biologists to detect gene activity. It hoped that the upgrade would have made the output files from its chips smaller and faster to process.

But researchers energetically protested that the new files would be incompatible with open-source software that has been developed by academics to work with the company's microarrays. Such software is free for everyone to use.

On 2 July, Affymetrix said that it would change course to ensure that its upgraded technology would not render the free software obsolete.

Microarray chips are covered in thousands of strands of DNA, each corresponding to a known gene and marked with a fluorescent molecular tag. Researchers add a sample of biological material to the chip, and the DNA in the sample bonds to the sequences on the chip. The chip outputs its results — a record of which dots are glowing, and how brightly — as a computer file. At present, these files can be opened on almost any computer.

Affymetrix announced late last year that it would upgrade its microarrays to produce



Chip panned: biologists prefer using open-source software to analyse their microarray data.

data as smaller files in a proprietary format. The change would have made it easier to modify formats in the future without disrupting the analysis software, says David Kulp, a bioinformaticist at the company.

But the move would have placed the microarray data files off-limits to the developers of open-source software. It would also have rendered useless many existing tools for interpreting the data, including a popular one called Bioconductor.

Many molecular biologists are happy with the simple file format produced by Bioconductor. "The file is readable by a human," says James MacDonald, who uses Bioconductor to analyse output from chips at the University of Michigan's microarray facility at Ann Arbor.

Researchers' worries were increased by the fact that US law now makes it difficult to write open-source software to interface with proprietary formats. On 27 June, 21 members of the group that develops Bioconductor announced that they would not develop software for a private file format.

"We wouldn't be willing to do anything that was marginally legal," says group member Robert Gentleman, a statistician at the Harvard School of Public Health. Researchers working on software in their spare time lack the resources to deal with proprietary file formats, he adds.

Besides the inconvenience to researchers forced to use new software, losing the efforts of a large group working out how best to analyse microarray data would have damaged the field as a whole, Gentleman argues. "The technology isn't mature yet. We need a lot of clever people working on these problems," he says.

Last week, after conferring with the Bioconductor team, Affymetrix announced that its new compressed format will be made public. "We don't have any trouble with this," says Kulp. "Clearly, some developers prefer to write their own file readers, which is more work on their part, but has advantages to open-source software distribution."

"I think that science won," says Gentleman. But he doubts that the issue will go away: "The clash between open source and commercialization will probably be a constant minefield for both sides." ■

UK shock tactics repel animal-rights activists in Japan

David Cyranoski, Tokyo

British animal-rights activists have pledged to export their fight to Japan. But animal-rights groups in Tokyo, after a brief dalliance with the outsiders, are recoiling from their aggressive tactics, which they say could backfire there.

Stop Huntingdon Animal Cruelty (SHAC), an animal-rights group based in Worcestershire, UK, said earlier this year that it planned to target Japanese academic and industrial researchers, saying that they make up one-fifth of the customer base of controversial UK animal-testing company Huntingdon Life Sciences. SHAC has already demonstrated against the European branches of many Japanese companies.

When SHAC member Dawn Hurst arrived in Tokyo in April to demonstrate at a pharmaceutical trade show, she was charged with the theft last year of a dog that was being used in experiments at Juntendo University. But after two directors of the

Tokyo-based Animal Rights Center (ARC) were also arrested in connection with the theft last week, the group has backed away fast from its association with SHAC.

Susumu Kawaguchi and Yukari Sugisaka of the ARC claim that they were keeping the dog at their shelter after it was given to them as a stray by Hurst, who has reportedly admitted stealing it. ARC representatives helped Hurst to find lodgings when she came to Japan but say that they were unaware of any illegal activity.

The ARC seeks to draw attention to animal cruelty in Japan. "Japanese consciousness of this problem is very thin," admits Chihiro Okada, another director of ARC. But Japanese sensibilities frown on the use of illegal means in the pursuit of political ends. "Given the Japanese character, radical activity would only alienate people," says Okada.

The ARC and larger animal-rights groups such as the Japan Antivivisection

Association (JAVA) say that they do not support SHAC's activities.

Hurst has also been charged with the theft of research materials relating to animal experiments in several Japanese laboratories. In August 2001, for example, she entered Osaka University's pharmacology department by posing as a graduate student from a British university, according to Osaka neuroscientist Masaya Tohyama. Hurst allegedly stole videos and photographs, which were later posted on SHAC's homepage.

Tohyama says that SHAC's tactics had some initial success, but that the laboratory was able to explain what it was doing and why. "The only thing that changed was that we became more wary of outsiders trying to get into the labs," he says.

But SHAC spokesman Greg Avery says that the group's activities have influenced the research of some Japanese companies, and that it plans to escalate operations against Japanese researchers. ■

Academic hopefuls set for pole position in US visa applications

Washington Academics and students could soon be able to jump the queue for US visas, thanks to new guidelines issued by the Department of State.

Since the terrorist attacks of 11 September 2001, applicants for visas have faced long security-related delays that some academic societies fear could drastically reduce enrolment (see *Nature* **423**, 906; 2003). In an attempt to prevent this, Janice Jacobs, the state department's deputy assistant secretary for visa services, asked consular officials last month to give priority to academics when setting up visa interviews.

"We're very pleased," says Barry Toiv, a spokesman for the American Association of Universities, which was informed of the development on 1 July. But he adds that US consulates still need more officials to help process the backlog that has built up since the attacks.



That sinking feeling: Japanese researchers lost the *Kaiko* submarine when its tether broke.

Ocean experts hit wave of bad luck as sub goes under

Tokyo Japanese researchers have given up the hunt for *Kaiko*, the world's deepest-diving submarine. The device has been missing since 30 May, when the tether connecting *Kaiko* to an underwater launcher snapped.

Kaiko had just finished collecting pressure data from holes drilled in the ocean floor 4.7 kilometres below the surface, south of the Kii peninsula in western Japan. Researchers from the Japan Marine Science and Technology Center had decided to reel in the craft, as an approaching typhoon was causing rough seas. The

tether snapped as they were bringing it back to the launcher.

Kaiko, which can dive to depths of 11 km, has carried out a wide range of missions since it was commissioned in 1995, from studying the Earth's structure to discovering bacteria that live in extreme deep-sea environments, and have potential medical use.

A spokesperson says that researchers hope to replace the ¥1.8-billion (US\$15-million) unit, but do not know whether they will be able to raise the money.

Test-tube babies 'not at risk of brain defects'

Paris The largest and longest-running study of babies born by assisted reproduction has eased fears that children conceived using the technology face an increased risk of psychological or cognitive problems.

The research, which was presented at the European Society for Human Reproduction and Embryology's annual meeting in Madrid last week, tracked around 1,000 children born by conventional *in vitro* fertilization or by direct injection of sperm into an egg, together with about 500 control children. Throughout a range of measures, no difference was found from birth until five years of age.

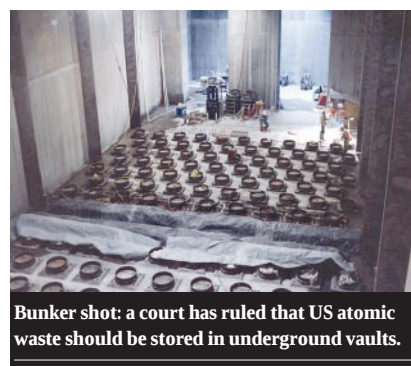
The authors say that the study was designed to detect cognitive problems, not physical defects. But measurements of physical development were taken, and children born using intracytoplasmic sperm injection, in which a sperm cell is injected into an egg, were found to have a higher rate of abnormalities such as urinogenital defects. The rate of defects for these babies was 6%, compared with 2% in control children.

Court rejects corner-cutting in bid to clear atomic waste

Washington A court judgement has overturned the US Department of Energy's masterplan to clean up sites contaminated with waste from its atomic-weapons programme. The planned endeavour is the world's largest environmental-restoration project.

The department's proposal, which would restore massively contaminated sites at Hanford in Washington state, Savannah River in South Carolina, and elsewhere, was revised in 1999 to prevent the project's costs — expected to exceed \$100 billion — from spiralling out of control. Under the revisions, some waste held in leaking underground tanks at these sites would have been kept *in situ*, instead of being turned to glass and stored permanently elsewhere.

But on 2 July, the Federal District Court in Boise, Idaho, ruled that the revised proposal



Bunker shot: a court has ruled that US atomic waste should be stored in underground vaults.

fails to comply with a 1982 law stipulating that highly radioactive waste should be stored permanently in underground chambers. The ruling is seen as a setback for the federal government, and as a victory for environmental groups and state governments, which have argued that more money should be spent on a thorough clean-up.

Oxford don penalized for outburst at Israeli

Tel Aviv Andrew Wilkie, a pathologist at the University of Oxford, UK, who rejected a PhD application because the student in question had served in the Israeli army, has been barred from selecting students and faculty members, pending the outcome of a disciplinary hearing.

The disciplinary committee will review an e-mail that Wilkie sent to the student, Amit Duvshani, who had enquired about the possibility of studying for a doctorate under Wilkie's direction (see *Nature* **424**, 7; 2003). Wilkie wrote that he would not accept a student who had served in the Israeli army because of his belief that Israel is committing human-rights abuses against Palestinians in the occupied territories.

He later issued a public apology, describing his message as a "wholly inappropriate expression of my personal opinions". An Oxford spokesman says that the university was "appalled that any member of its staff should have responded to an enquiry from a potential graduate student" in the terms that Wilkie used.

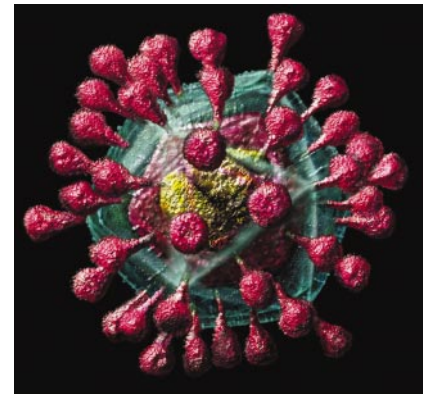
Correction: Broad Institute

In the news in brief "Whitehead geneticist quits for broad remit at medical institute" (*Nature* **423**, 910; 2003) it was incorrectly stated that Eric Lander was leaving his post as director of the Whitehead Institute at the Massachusetts Institute of Technology (MIT) to move to the Broad Institute in Cambridge, Massachusetts. Lander is director of the Whitehead Institute/MIT Center for Genome Research. This centre will become a part of the Broad Institute and Lander will remain in his position as director.

SARS

What have we learned?

It's less than four months since the World Health Organization issued global warnings about a mysterious and deadly form of pneumonia. *Nature's* reporters pose key questions about the outbreak, and assess our preparedness to deal with future viral threats.



R. KIGHTLEY/SPL



K. CHEUNG/REUTERS

High alert: caused by a coronavirus (top), SARS left many in Hong Kong in the grip of fear.

Was the fuss overblown?

To the friends and relatives of the 800-plus people slain by severe acute respiratory syndrome (SARS), this might seem like a callous question. But SARS barely registers a blip in the annual body count caused by infectious disease. Influenza is likely to kill up to half-a-million people in 2003, whereas the death tolls from malaria, tuberculosis and AIDS will each run to seven figures. So, when the outbreak is put into perspective, was the panic over SARS really warranted?

Given the economic damage suffered by those countries named by the World Health Organization (WHO) in its warnings against travel to the worst affected regions, some commentators have accused the WHO of overreacting—local officials in the Canadian city of Toronto, in particular, were livid about the agency's advice against travelling there. But the experts contacted by *Nature* are unanimous in rejecting the general idea that health authorities went over the top.

When reports of an unusual respiratory illness began to emerge from southern China

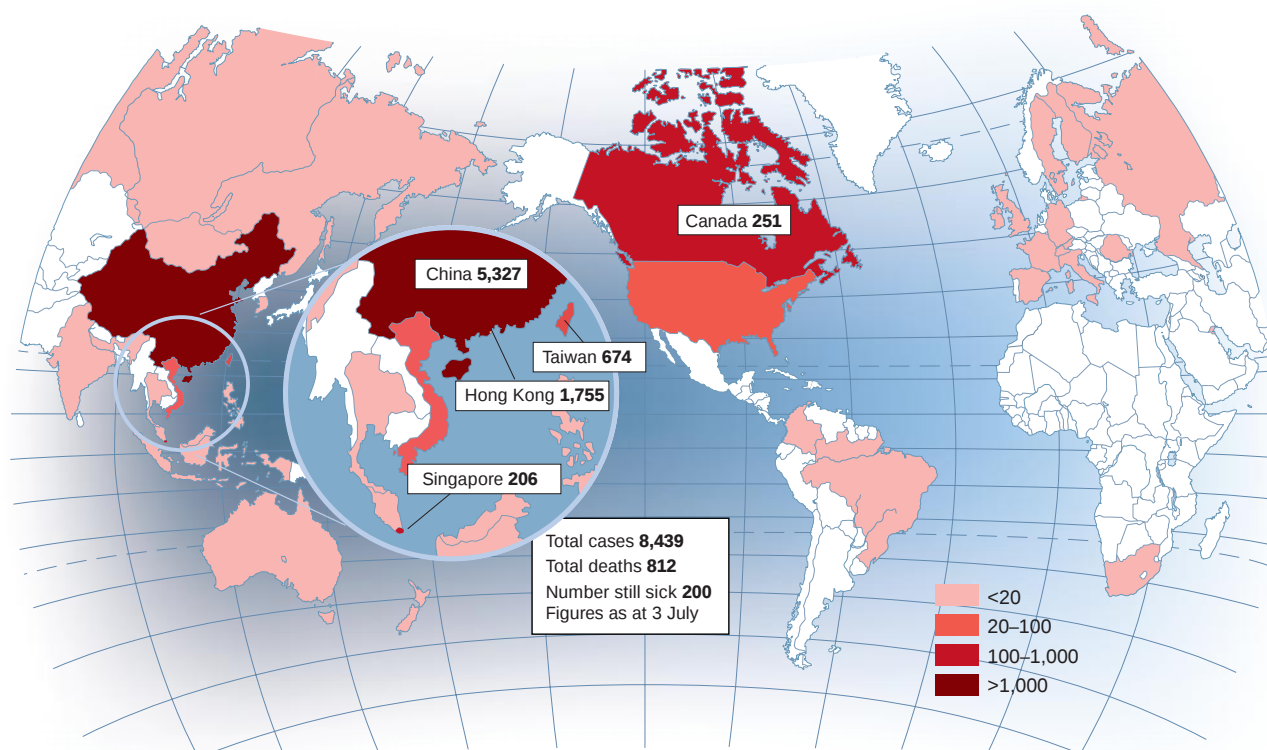
late last year, no one knew what caused it. Was this a new super-virulent strain of flu? If so, would it be as deadly as the 1918 pandemic strain, which killed up to 40 million people? Even in mid-March, when outbreaks elsewhere in Asia caused the WHO to release the two global alerts that shot SARS onto the world's news agenda, the culprit remained unknown. Also unclear were the disease's death rate and infectiousness. Health officials could not be sure whether they were dealing with a troubling but ultimately limited threat, or a global mass killer.

Thankfully, it seems that SARS isn't sufficiently infectious to cause a re-run of the 1918 flu pandemic. Even so, the relatively low death count can probably be attributed in large part to the surveillance and patient isolation rapidly introduced in most of the countries that imported the disease. China has been widely criticized for the initially sluggish and secretive response that allowed SARS to take hold within its borders. But after the WHO issued its global alerts, only Taiwan—where officials failed to rapidly establish a central coordinating office for SARS—experienced an outbreak in which there was widespread

transmission among people other than health workers caring for SARS victims.

Although experts agree that SARS warranted a vigorous reaction, questions remain about the way in which the threat was communicated to the public. The WHO's executive director for communicable diseases, David Heymann, acknowledges that problems were caused by the agency's second global alert, issued on Saturday 15 March when some ministries of health were closed, after evidence emerged that the disease was being spread internationally. As a result, the media were running with the story before national officials had been briefed about how to respond. To rectify this, the WHO's general assembly in May urged member states to designate officials who will be available around the clock to be informed about future urgent global alerts.

Another problem was that some national agencies, even the respected US Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia, released information that inadvertently exaggerated the ease with which SARS spreads. For example, in a editorial in *The New England Journal of Medicine*¹



Pattern of an epidemic: the total number of SARS cases worldwide (above) was contained thanks to global alerts and patient quarantine; and the number of new cases (right) has now tailed off.

published online on 2 April, CDC director Julie Gerberding wrote: "Airborne transmission may have a role in some settings." No wonder that people in Hong Kong and elsewhere were scared into wearing surgical masks in the street, even though clinical reports suggested that SARS was being spread only by close personal contact.

But these were minor glitches. Experts agree that it is better to be accused of overreacting than of allowing the disease to run out of control. If a similar fuss had erupted in the early days of AIDS, suggests epidemiologist Megan Murray of the Harvard School of Public Health in Boston, maybe HIV would not now be killing three million people each year.

Helen Pearson

Is the outbreak finally over?

With luck, the answer is yes. At the outbreak's peak in early May, some 200 new cases were being reported every day. But as *Nature* went to press, no new cases had been reported since 15 June. "In some senses, we should breathe a sigh of relief," says Roy Anderson, whose team at Imperial College London has been at the forefront of efforts to characterize the spread of the disease.

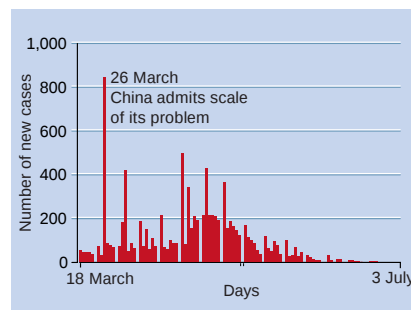
For Anderson and other epidemiologists who have been crunching the numbers on SARS, the key parameter is known as R_0 . A measure of a disease's infectiousness, R_0

corresponds to how many people, on average, are infected by each patient in the absence of any control measures.

Attempts at modelling the spread of SARS, published online on 23 May by teams led by Anderson and by Marc Lipsitch of the Harvard School of Public Health^{2,3}, gave R_0 a value of between two and four. This was encouraging news, as it confirmed that efforts to isolate patients should bring the disease under control. Contrast this with flu, which boasts an R_0 of about 10. For diseases this infectious, quarantining those who show symptoms is not enough to bring the average number of new infections caused by each case to below one — the level necessary for an epidemic to go into decline.

But we still know too little about SARS to predict how it will behave should it return. Key unknowns include how quickly a person becomes infectious after they are infected, and how long they remain able to spread the disease. Most importantly, we don't know exactly how the virus is transmitted, how many people harbour the SARS virus without showing symptoms, and whether these 'silent' cases can infect others.

The relatively sluggish transmission of SARS fits with the idea that it is spread by close contact, requiring its victims to breathe in droplets of virus-laden mucus. That explains why most transmission has been in confined settings, such as hospitals. But unusual foci of infection in Hong Kong show



that this isn't the whole story. Most striking was the outbreak at the Amoy Gardens apartment block, where the brother of one resident seems to have infected 321 other people. Scientists tore the block apart, eventually implicating a faulty sewage system that allowed droplets contaminated with faeces to form in the block's bathrooms⁴. Transpose SARS to a developing country with poor sanitation, therefore, and the dynamics of its spread might be different.

If SARS does return at a later date, its epidemiology could be different. For reasons that aren't fully understood, respiratory infections spread more rapidly during the winter. Another crucial question is how long immunity to SARS persists. If this is as short as a few months, and SARS bounces back with the next Northern Hemisphere winter, even those previously exposed to the virus may be just as vulnerable as they were the first time around. "It's quite possible that we haven't seen SARS at its full force," warns Donald Burke, an international health

expert at Johns Hopkins University in Baltimore, Maryland.

In the meantime, epidemiologists are trying to work out how many people are silently harbouring the SARS virus. If a sizeable number of symptomless cases can transmit the disease, outbreaks will continue to erupt. Investigating this question will require better diagnostic tests — to detect both antibodies to the SARS virus, and the virus's genetic material. Ideally, these won't require laborious analysis of blood samples. "We need a saliva test," says Anderson.

The greatest danger is if health officials let down their guard now that SARS seems to have waned. A second outbreak in Toronto in late May, weeks after the last new case was reported in the city, provided a warning. In this instance, a patient was discharged from hospital with pneumonia — common among elderly people who have undergone surgery — that turned out to be SARS.

Tom Clarke

Are we prepared for the next viral threat?

SARS should be seen as a warning shot. If the virus had been more infectious, we could now be facing millions of deaths. And this nightmare vision doesn't require the emergence of an entirely novel disease: new strains of flu virus arise each year, and every few decades, one appears that wreaks global havoc. So, in the light of our experience with SARS, are we prepared? Scientifically, the answer is a qualified 'yes'. In terms of public health, it's a resounding 'no'.

The most encouraging thing about the response to SARS is the way in which virologists worked together to identify and understand the pathogen responsible. The four collaborating centres in the WHO's Global Influenza Surveillance Network — in Australia, Britain, Japan and the United States — were soon teamed up with seven other leading virology laboratories to probe the mysteries of SARS. Within a few weeks of the WHO issuing its global health alerts, labs in this network had identified the culprit as a previously unidentified strain of coronavirus⁵⁻⁷, and a prototype diagnostic test was made available⁷.

The network was all about sharing data, resources and time. Results were posted on a



The dark lesions on these X-rays show the debris and fluid blocking the lungs of patients with SARS.

password-protected website as soon as they were gathered, and were discussed daily in teleconferences that sometimes stretched over three hours. Most helpful, say insiders, was the prompt reporting, from the word go, of blind alleys. "We had already run through the list of obvious possibilities when the network started, and so could tell the partners not to look for certain viruses," says network member John Tam of the Chinese University of Hong Kong.

Sharing of materials and reagents was also fundamental. The SARS agent was shown to be a coronavirus by amplifying portions of its genetic material using the polymerase chain reaction (PCR), and examining the resulting sequences. But this approach is only feasible if you have the appropriate PCR 'primers', corresponding to short genetic sequences from known virus families. Without the sharing of primers across the collaborating network, says William Bellini, SARS laboratory coordinator at the CDC, the virus would not have been identified so quickly.

WHO officials attribute the rapid scientific progress made by the SARS network in part to the experience of the influenza labs around which it was constructed. "We had a core of people, definitely not ivory-tower people, who were used to working in this way," says Klaus Stöhr, the WHO's influenza project leader, and its resident SARS expert. He is now building further networks to share clinical experience and data from SARS patients, and to link groups trying to identify animal reservoirs of the SARS virus.

Unfortunately, a speedy scientific reaction to an emerging viral threat doesn't guarantee that public-health officials will be able to mount an effective response. Faced

with a more rapidly spreading disease, the consensus is that they would have been overwhelmed. And although health authorities might claim that it is unreasonable to expect them instantly to be able to tackle a disease that pops up out of the blue, most nations remain woefully unprepared to deal with a flu pandemic that could emerge at any time.

Through the WHO's efforts, emerging strains of flu virus are constantly monitored in the hope that vaccines can be produced that protect against those in general circulation. But if a pandemic strain does emerge, health officials will also need to deploy anti-flu drugs on a massive scale. "We know that another flu pandemic will occur," says Albert Osterhaus, a virologist at Erasmus University in Rotterdam, the Netherlands, who participated in the SARS network. "But no country has yet started to stockpile antiviral drugs."

Alison Abbott

How and who does SARS kill?

No two cases of SARS are exactly the same. Depending on the age and fitness of the patient, the disease can run wildly different courses. Even the symptoms of fever and dry cough, initially included in the case definition for SARS, are no longer considered to be universal.

One pivotal point seems to occur at about the beginning of the third week after infection, when some patients, especially the young, improve. Others, however, progress to a more severe form of the disease — their lungs become clogged with debris and fluid, which show up as dark lesions in chest X-rays. In about a fifth of all patients, this

SARS facts

- **Symptoms** Initially fever, sometimes accompanied by headache; later a dry cough, trouble with breathing, nausea and diarrhoea.
- **Incubation period** 2–10 days
- **Economic cost** Nearly US\$100 billion, mostly as a result of cancelled travel and decreased investment in southeast Asia, according to Bio Economic Research Associates of Cambridge, Massachusetts.

requires aggressive treatment such as mechanical ventilation. Even then, many of these people die.

Worldwide, the death rate from SARS seems to be about 10%. But individual risk factors vary considerably. For people over 65 years of age, more than half of those infected will die. Just about any lung ailment complicates the disease, and conditions such as emphysema are more common in the elderly. Other concurrent infections may also be involved. Although it is now well established that the SARS virus can kill on its own⁸, other viruses that have been isolated from patients with SARS⁹ could exacerbate the illness.

The ultimate cause of death also remains unclear. Does the virus kill directly by destroying cells in the lung, or does the immune system deliver a *coup de grâce* by fighting back too hard? By the time that most of the lung damage occurs, the amount of virus circulating in the blood has already peaked¹⁰, suggesting the latter. And the pattern of damage is consistent with an overload of cytokines⁹ — biochemical messengers that rev up our immune responses. But for the time being, pathologists are recording an open verdict.

Jonathan Knight

Where did the SARS virus come from?

The SARS coronavirus is believed to have jumped over from an animal host to people in rural areas of Guangdong province in southern China. From November last year, it circulated there for several months while Chinese health authorities failed both to tackle its spread, and to provide adequate information to their counterparts in other countries about what was going on. But the path that the virus took to set up this initial hotbed of human infection — essential information for assessing the likelihood of a recurrence, even if the initial wave of SARS is over — remains a mystery.

Coronaviruses are named after their crown-like halo of protein spikes, which help them to latch on to their host cells. Those previously identified in people cause nothing nastier than common colds, but some of the coronaviruses that afflict livestock and pets cause more serious conditions.

Analysis of the complete genome sequence of the SARS virus, published in May¹¹, suggests that it is not closely related to any of the three previously identified coronavirus subfamilies, nor does it seem to have arisen through a chance genetic recombination between known coronaviruses¹². “Its unique sequence suggests that it has evolved independently from the other members of the family, in some animal host, for a long time,” says Malik Peiris, a virologist at the University of Hong Kong.

Ongoing research by Peiris and his col-



Wild animals, such as the palm civet, sold in Chinese markets, might have passed SARS to humans.

leagues may shed light on the origins of the virus. The Hong Kong team is looking at genomic sequences of coronaviruses sampled from masked palm civets (*Paguma larvata*) and other animals sold in the markets of southern China. Comparison of the sequences of the viruses found in different animals should make it possible to trace the evolution of the SARS virus and determine which animal passed the disease to humans. Yi Guan, another member of the Hong Kong team, says that related viruses have so far been found in about half-a-dozen species — which he declines to name until the work has been published.

Knowledge of the chain of animals involved in passing the SARS virus to humans would help in the design of preventive measures. For example, when the previously unidentified Nipah virus began causing fatal encephalitis in livestock and people in Malaysia in 1998, about one million pigs were slaughtered. Later the

virus was found to reside in fruit bats¹³, so farmers could take measures to isolate their livestock from this natural reservoir. “Once you find the source, you can find out how to manage it better,” says John Mackenzie, a virologist at the University of Queensland in Brisbane, Australia.

It will probably be some time before we pin down the natural reservoir for SARS. Recent investigations by researchers at the China Agricultural University in Beijing, for instance, have failed to find SARS-like coronaviruses in 732 animals from 54 wild and 11 domestic species in southern China, including palm civets. As with efforts to investigate the epidemiology of SARS in people, progress may depend on the development of improved diagnostic tests. But potentially, Guan warns, revealing the origins of SARS could require decades of painstaking fieldwork.

David Cyranoski

Why China?

SARS is not the first viral disease to burst out of China or Hong Kong. The southern Chinese region was the source of influenza pandemics in 1957 and 1968, and scares about the transmission to people of novel strains of avian flu in 1997 and 2001. Why does this region keep throwing up viruses that have the potential to threaten the lives of people around the world?

Southern China's status as the world's primary breeding ground for new strains of flu is explained by the fact that its people, pigs and domestic fowl, which all harbour influenza viruses, live cheek-by-jowl, increasing the likelihood that two strains will recombine genetically to produce a deadly new variant. “The animals walk in



Protect and survive: residents of Beijing make their way to work at the height of the SARS crisis.



Despite the propaganda, China has been criticized for its initially slow and secretive response to SARS.

and out of their houses," says Kenneth Shortridge, who led the University of Hong Kong's efforts to monitor avian viruses in southern China until his retirement last year. Preliminary evidence suggests that SARS followed a different model, apparently crossing over to people from wild animals, rather than livestock. But this, too, is not terribly surprising, given that the southern Chinese make widespread use of wild species for food and traditional medicine — practices that Chinese health officials are now trying to discourage.

Another dietary issue — specific nutritional deficiency — has also been tentatively linked to the emergence of new viral strains in rural China. For instance, in many parts of the country, the diet is lacking in the trace element selenium. A team led by Melinda Beck of the University of North Carolina at Chapel Hill found that when the coxsackievirus B3 infects mice deficient in selenium, it mutates at a much higher rate and can become more virulent¹⁴. Beck suspects that this phenomenon may explain the high incidence of Keshan disease, a weakening of the heart muscle, in some Chinese populations¹⁵. She has also observed increased mutation rates in flu viruses infecting selenium-deficient mice¹⁶. "The fact that China has widespread selenium-deficient areas may play a role in the emergence of new viral strains," Beck claims.

Other scientists regard Beck's findings as speculative, and doubt whether they offer a general explanation for the emergence of viral diseases in China. When you have the world's largest population interacting closely with livestock and wild animals, say experts, it's hardly surprising that China seems to be the origin of so many viral outbreaks. "It's a

matter of exposure probability," suggests Mei-Shang Ho, an epidemiologist with Academia Sinica's Institute of Biomedical Sciences in Taipei, Taiwan.

David Cyranoski

Is the SARS virus mutating?

Viruses such as HIV and those that cause influenza have often been described as 'wily' because they mutate rapidly, a trait that helps them to evade drugs or the human immune system. But so far, the SARS virus seems remarkably invariant: the genome sequences of 14 isolates from patients in Singapore, Toronto, China and Hong Kong have not revealed any changes of real consequence¹⁷.

This isn't because the SARS virus fails to mutate, but rather that the mutations thrown up so far haven't proved to be particularly beneficial to it. As the virus has so far encountered little resistance from its new human hosts, there has been little selective pressure to cause new mutants to be retained.

Coronaviruses are quite sloppy when it comes to replicating their genetic material, making one error for every 10,000 nucleotides they copy — roughly the same error rate as HIV. But coronaviruses have a trait that allows them to weed out mutations as they occur. Rather than relying on a single template genome, the enzyme responsible for copying the viruses' genetic material sometimes jumps around between multiple copies of the viral genome present in an infected cell. So each new genome is actually copied from several templates, reducing the chance that any given mutation will become entrenched in the viral population.

But if one of these jumps is imprecise, a whole chunk of genome can get skipped, resulting in the deletion of part of an important gene. The consequences can be dramatic, particularly if the change affects the protein spikes that bind to the surface of the viruses' cellular victims. For example, in 1984 a new respiratory ailment appeared on European pig farms. It turned out to be a deletion mutant of a coronavirus that previously had infected piglets' stomachs¹⁸. The altered spike protein had changed the type of cells the virus could enter. Although the new disease was not generally lethal, it has since spread worldwide and complicated diagnosis of the gut disease.

A genetic deletion may also have helped the SARS virus to make the transition from its animal reservoir to humans. But, if so, it is a different type of change — the spike protein remains intact. Instead, compared with the viral strains found in animals on sale in southern Chinese markets, the SARS virus lacks 29 nucleotides in the gene for a protein of unknown function, which is attached to the inside of the virus's protective coat.

Should SARS return to haunt us, it will probably not remain as stable as it has been so far, particularly if it is attacked with antiviral drugs. Our immune systems could force changes, too. "Once enough people develop immunity, mutations will be favoured, just as you see with flu viruses," predicts Michael Lai, a molecular virologist at the University of Southern California in Los Angeles.

Jonathan Knight

Are drugs for SARS on the horizon?

From the moment the viral culprit behind SARS was unmasked, drug-discovery researchers leapt into action. So far, the main approach has been one of brute force: screening hundreds of thousands of compounds for their ability to attack lab cultures of the virus.

The US National Institute of Allergy and Infectious Diseases in Bethesda, Maryland, is coordinating a massive random screen of both licensed drugs and those still under development. This work has been contracted to the US Army Medical Research Institute of Infectious Diseases at Fort Detrick, also in Maryland, where more than 300,000 compounds — many of them supplied by pharmaceutical companies — have so far been tested on viral cultures grown in monkey kidney cell lines. "We've had lots of hits, and some are looking better than others," says Fort Detrick virologist Robert Baker.

A similar, but smaller initiative, based at the University of Frankfurt in Germany, has shown that a compound called glycyrrhizin, derived from liquorice roots, can rid monkey kidney cells of the SARS virus¹⁹. In work that has yet to be published, the Frankfurt team has

confirmed the effectiveness of glycyrrhizin in a human cell line. Although relatively non-toxic and already licensed for use in conditions including hepatitis C, glycyrrhizin only works at very high doses. So the Frankfurt researchers, led by Prakash Chandra, are collaborating with medicinal chemists at the Russian Academy of Sciences' Institute of Organic Chemistry in Moscow, who have synthesized a series of related compounds²⁰. They hope that one of these will prove particularly effective against the SARS virus.

Other researchers are trying a more directed approach. Erik De Clercq of the Catholic University of Leuven in Belgium, for instance, is screening selected compounds from his large library of antiviral chemicals, many of which interfere with viral replication. "I believe that rational screening based on putative targets is likely to be more efficient than random screens," he says.

Rolf Hilgenfeld, a structural biologist at the University of Lübeck in Germany, meanwhile, has solved the structure of a key SARS enzyme called proteinase, which turns viral proteins into the active forms required for viral replication²¹. Using a computer model of this structure, his team has also begun to predict which drugs might inhibit the enzyme's activity. Hilgenfeld is now collaborating with a Chinese group led by Jiang Hua-Liang of the Shanghai Institute of Materia Medica, which has the supercomputing power to expand upon this work.

But the difficult part will be moving into animal experiments and eventual human trials. So far, there is only one validated animal model for SARS, the cynomolgus macaque (*Macaca fascicularis*)⁸, which isn't ideally suited for large-scale investigations of candidate drugs. A good small-animal model is urgently needed, say researchers. **Alison Abbott**

What about a vaccine?

If SARS stages a comeback, the best tool for blunting its threat will be an effective vaccine. And the good news is that vaccines already exist for animal coronaviruses. "We can immediately apply this expertise to SARS," says virologist Peter Rottier of Utrecht University in the Netherlands, who is developing a vaccine against a coronavirus that kills cats. Another encouraging sign is that the condition of SARS patients seems to improve if they are given serum from previously infected people, which indicates that human antibodies can neutralize the virus.

Perhaps the easiest approach is to stimulate immunity using a killed SARS virus. "It's the first thing we'll try," says Rino Rappuoli of vaccine manufacturer Chiron in Siena, Italy. But relying on killed viruses is not ideal — in part because ensuring that all viruses in a vaccine are dead and yet retain the ability to stimulate the immune system is tricky.

The next option is a weakened SARS virus



Fighting back: serum from previously infected people seems to help treat other SARS patients.

that can survive in humans long enough to challenge the immune system, but which doesn't cause disease. Such vaccines are normally made by culturing viruses in animal cell lines for generation after generation, selecting each time for the least potent offspring. They have the advantage that they can be made to infect cells in the respiratory tract — which may prove crucial to stopping SARS in its tracks. But safety remains an issue, as a weakened strain might mutate to become a lethal virus in its own right.

The best way around this, and the approach that Rottier has used to make a prototype cat coronavirus vaccine, is precisely targeted genetic modifications. Genes not needed for the virus's survival but required for it to cause disease are removed. All of the known coronaviruses, including the SARS virus, seem to have these genes in common, and knocking them out wholesale would make it almost impossible for the SARS virus to mutate back to a dangerous form. But there is still the chance that it could recombine with other coronaviruses to recover its lethal genetic machinery.

Other approaches avoid any possibility of a vaccine causing SARS. For instance, harmless viruses could be engineered to contain genetic sequences from the SARS virus. Such vaccines could again be made to infect cells in the respiratory tract, and the approach has been used successfully in animals — for instance in a prototype vaccine against a coronavirus that causes bronchitis in chickens²². A simpler and even safer alternative would be vaccines based on viral proteins that stimulate the immune system, but this approach has had only limited success against animal coronaviruses.

What works in animal diseases may not provide a perfect guide to developing a SARS

vaccine, however. Experience has shown that individual coronaviruses can interact with their hosts in quite distinct ways. "The trick is finding a vaccine that pushes all the right immunological buttons," says Dave Cavanagh at the Institute for Animal Health at Compton in Berkshire, UK. Finding a candidate that achieves this against the SARS virus will require extensive studies in animals.

If all goes well, a SARS vaccine could reach the market in as little as four years, say experts. But there's a lot that could go wrong at any stage. So for the foreseeable future, health officials had better plan on tackling the disease without this key defensive weapon. **Tom Clarke**

1. Gerberding, J. L. *N. Engl. J. Med.* **348**, 2030–2031 (2003).
2. Riley, S. *et al. Science* **300**, 1961–1966 (2003).
3. Lipsitch, M. *et al. Science* **300**, 1966–1970 (2003).
4. http://www.info.gov.hk/info/ap/pdf/amoy_e.pdf
5. Peiris, J. S. M. *et al. Lancet* **361**, 1319–1325 (2003).
6. Ksiazek, T. G. *N. Engl. J. Med.* **348**, 1953–1966 (2003).
7. Drosten, C. *N. Engl. J. Med.* **348**, 1967–1976 (2003).
8. Fouchier, R. A. M. *et al. Nature* **423**, 240 (2003).
9. Nicholls, J. M. *et al. Lancet* **361**, 1773–1778 (2003).
10. Peiris, J. S. *et al. Lancet* **361**, 1767–1772 (2003).
11. Marra, M. A. *et al. Science* **300**, 1399–1404 (2003).
12. Rota, P. A. *et al. Science* **300**, 1394–1399 (2003).
13. Yob, J. M. *et al. Emerg. Infect. Dis.* **7**, 439–441 (2001).
14. Beck, M. A., Shi, Q., Morris, V. C. & Levander, O. A. *Nature Med.* **1**, 433–436 (1995).
15. Levander, O. A. & Beck, M. A. *Biol. Trace Elem. Res.* **56**, 5–21 (1997).
16. Nelson, H. K. *et al. FASEB J.* **15**, 1846–1848 (2001).
17. Ruan, Y. *et al. Lancet* **361**, 1779–1785 (2003).
18. Pensaert, M., Callebaut, P. & Vergote, J. *Vet. Q.* **8**, 257–261 (1986).
19. Cinatl, J. *et al. Lancet* **361**, 2045–2046 (2003).
20. Baltina, L. A. *Curr. Med. Chem.* **10**, 155–171 (2003).
21. Anand, K., Ziebuhr, J., Wadhwani, P., Mesters, J. R. & Hilgenfeld, R. *Science* **300**, 1763–1767 (2003).
22. Johnson, M. A., Pooley, C., Ignjatovic, J. & Tyack, S. G. *Vaccine* **21**, 2730–2736 (2003).

♦ www.who.int/csr/sars/en/index.html

♦ www.cdc.gov/nccid/sars

♦ www.sars.gc.ca

♦ www.info.gov.hk/info/sars/eindex.htm

Nature's SARS focus

♦ www.nature.com/nature/focus/sars

History and science united to vindicate Perutz

A generous team player, he was backed by colleagues in clearing his name over DNA.

Sir — Floods of ink, electronic and actual, have been spent on celebrating the fiftieth anniversary of the discovery of DNA's structure, recollecting facts and testimonies from the protagonists (see www.nature.com/nature/DNA50). A gripping correspondence that I recently uncovered among the papers acquired by Jeremy Norman in Novato, California, adds a further motif of reflection on one much-debated episode.

The facts are well known. André Lwoff of the Pasteur Institute in Paris, reviewing Jim Watson's book *The Double Helix* (Athenaeum, New York; 1968), highlighted a crucial sentence that implied a breach of faith by Max Perutz, who seemed to have given key scientific material to his colleagues Watson and Crick (*Scientific American* **219**, 133–137; 1968), without permission. Soon after, biochemist Erwin Chargaff denounced Perutz for the same reason (*Science* **159**, 1448–1449; 1968).

The precious information was part of a Medical Research Council (MRC) report that Sir John Randall, director of King's College London, had circulated to all the members of the Biophysics Research Committee, including Perutz, who visited his unit on 15 December 1952. As Watson and Crick both acknowledged, the report, containing Rosalind Franklin's precise measurements, gave them an important clue that helped them unravel the anti-parallel nature of DNA and scoop both the

London team and Linus Pauling of Caltech.

A year later, Perutz published a rebuttal to Lwoff and Chargaff, in which he described the non-confidential nature of the MRC report (*Scientific American* **221**, 8; 1969 and *Science* **164**, 1537–1538; 1969). Why did it take him so long to respond?

What was not known until now is that Perutz's brief rebuttals were the result of backstage scientific teamwork starting in 1968. This correspondence, which I have now uncovered, offers a fascinating look at a chorus of scientists determined to set an official version of events before the public, and gives an insider perspective on the politics of scientific discovery.

"As you may have gathered," Perutz wrote to a colleague, "I am anxious to correct a story in Watson's book which suggests that I gave Watson and Crick a confidential report... I have now drawn up the enclosed letter which I propose to send to the *Scientific American* and to *Science*... Randall, Wilkins, Gosling and Crick have all agreed. So have Lwoff... and Chargaff. ... Watson is writing an addition to my letter, with an apology for having misrepresented the incident."

Perutz, who died early last year (see *Nature* **415**, 851–852; 2002), was an extraordinary experimentalist, a hard worker at the bench and a cautious man of science. Not interested in the rat race, he just wanted to solve problems. This time, the

issue was to prove his innocence. "Watson asked ... why I had not shown the report to [Watson] and Crick immediately," he wrote to Randall. "Was it because I did not want them to get on with the model, etc? And if the report was not confidential, then surely there was no need to ask your permission to show it to them."

Perutz went into the historical evidence, scrupulously interviewing his colleagues and MRC administrators while going through his initial draft, sentence by sentence, to build up a definitive version.

Two letters in particular are worth mentioning. Randall's prompt reaction to being asked whether it was wise to open the lid was to firmly advise leaving it to the historians. Maurice Wilkins, however, took the opposite view, because he thought that science and its discoveries belong to scientists, in the sense that they are the only ones who can fully understand them and therefore say what the true facts are.

To quote the biochemist John Edsall, presenting Perutz's letter to the editor of *Science*: "An important issue of scientific ethics is involved and the matter should be cleared up not only in order to do justice to Perutz but also because of the influence that this episode may have on other people."

Marta Paterlini

Laboratory of Human Neurogenetics,
The Rockefeller University, 1230 York Avenue,
New York, New York 10021, USA

Were Cro-Magnons too like us for DNA to tell?

Sir — Your News report "Anthropologists cast doubt on human DNA evidence" (*Nature* **423**, 468; 2003) refers to our study of two 24,000-year-old anatomically modern Europeans, whose mitochondrial DNA sequences were similar to those in today's humans (D. Caramelli *et al.* *Proc. Natl Acad. Sci. USA* **100**, 6593–6597; 2003). You report views from the community that ancient DNA evidence cannot be used to draw conclusions about the evolution of modern humans because of fundamental problems in the technique.

The criticisms you report are not based on the quality of our study. Neither of the quoted specialists questions our methodology, which was defined by one of them (A. Cooper and H. N. Poinar, *Science* **289**, 1139; 2000). The problem is in the results.

None of the nine tests we carried out suggests contamination by modern DNA, but the ancient sequences we determined

look modern, so there is a suspicion that they are indeed modern, the result of contamination. You report views that the DNA of Cro-Magnons can be studied only if they were different from us. If they had the bad luck to be like us, their sequences must remain unknown forever.

That is an unusual way to conceive science, and one that leads to paradox. If we are to apply this criterion of certainty to other areas, we should have abandoned anaesthesia (it may have side effects), air transportation (planes can fall down), cooked food (it may burn your fingers) and sexual reproduction (you might get AIDS).

If it can be shown that there is an error in our paper, we shall be happy to reconsider our conclusions. Should our results be confirmed, we will be even happier. But if we did all the right things, as seems to be the case so far, it seems irrational to question our study just because one never knows.

Guido Barbujani, Giorgio Bertorelle

Department of Biology, University of Ferrara,
Via L. Borsari 46, 44100 Ferrara, Italy

Multitude of reference styles delays publication

Sir — Peter Kareiva *et al.* make a compelling argument in Correspondence (*Nature* **420**, 15; 2002) that "slow-moving journals hinder conservation efforts". Our appraisal of the 14 journals they assessed shows that time in review is at least partly related to length of reference sections, and citation lists in the three conservation and applied-ecology journals are among the longest. Although many factors affect speed of publication, surely time lost by scientists and scientist-editors on formatting journal-specific reference sections is one of the most frustrating.

There may not be as many reference styles as there are journals, but sometimes this can appear to be the case to a beleaguered author. The considerable time wasted on formatting citation lists into a journal's style is exacerbated by the high likelihood of a submitted paper being rejected and requiring resubmission

elsewhere, perhaps more than once. The bibliographic software developed to assist with formatting citations only encourages the problem, in that individuals and institutions are compelled to spend their limited money on products that require scientists to spend time learning software rather than doing research. And, of course, the software does not encourage progress toward standardization.

A common citation format, familiar to all, would be sensible. It has been achieved for online versions of papers with the Digital Object Identifier system (DOI), so it should be possible for print journals. Coordination could be undertaken by an organization such as the International Council for Science via its committee on dissemination of scientific information, or the Society for Scholarly Publishing. Certainly, the substance of papers in journals is more important than uniquely stylized references. Time is precious to scientific pursuit, and time wasted retards discovery.

David M. Leslie, Jr*, Meredith J. Hamilton†

**US Geological Survey, Oklahoma Cooperative Fish and Wildlife Research Unit, 404 Life Sciences West, Oklahoma State University, Stillwater, Oklahoma 74078, USA*

†Department of Zoology, Oklahoma State University, Stillwater, Oklahoma 74078, USA

Writing: visuals are another story

Sir — Your News Feature “Clear as mud”¹ highlights an important issue. The course on effective writing that we run for PhD students here has taught us that instruction on elements of graphic design (one of us, K.M.F., is a medical illustrator) complements very well the linguistic aspects you address.

We see writing a paper as a business of integrating written and visual information. Scientists are prone to create over-elaborate graphs, charts and other illustrations which inadequately highlight key data. This may actually make a good text more difficult to read. Even when a lot of data need to be included, the manageable information content needs to be taken into account by the author.

For instance, a table containing 17 columns and 7 rows provides necessary documentation but has little visual impact (see, for example, Table 2 of ref. 2). Such tables have greater impact when they are large enough to allow balanced inclusion of text and numerical data (as in Table 1 of ref. 3). Also, complex tables can be made clearer to read by descriptions using full words rather than numerous abbreviations. When experimental results are presented, panels showing single comparisons

between control and experimental groups (as in Fig. 4 of ref. 4) have a visual advantage over graphs that show multiple experiments on the same axis (see Fig. 3 of ref. 5).

Scientific journals often insist on a very traditional, bland presentation of text and graphics. Perhaps a more creative approach to this on the part of editors would benefit writers and readers alike.

To summarize, before an author submits a paper to a journal, he or she should ask what kind of visual impact it has. Attention to this question would not only improve the paper's readability but would also allow referees to concentrate on data without getting frustrated by the lack of clarity. It might even decrease the rejection rate of essentially good papers.

Marek H. Dominiczak, Kathy McFall

Medical Humanities Unit, Gartnavel General Hospital, 1053 Great Western Road, Glasgow G12 0YN, UK

1. Knight, J. *Nature* **423**, 376–378 (2003).
2. Harborne, L., Fleming, R., Lyall, H., Norman, J. & Sattar, N. *Lancet* **361**, 1894–1901 (2003).
3. Advanced Bladder Cancer (ABC) Meta-analysis Collaboration *Lancet* **361**, 1927–1934 (2003).
4. Golder, M. et al. *Lancet* **361**, 1945–1951 (2003).
5. Sobngwi, E. et al. *Lancet* **361**, 1861–1865 (2003).

Writing: LEX and flexibility

Sir — In your News Feature on unclear writing, “Clear as mud” (*Nature* **423**, 376–378; 2003), a passage is cited from the abstract of a paper of which I am first author: “A somitic compartment of tendon progenitors” (*Cell* **113**, 235–248; 2003). The passage reads: “We demonstrate that the tendons associated with the axial skeleton derive from a heretofore unappreciated, fourth compartment of the somites. *Scleraxis* (*Scx*), a bHLH transcription factor, marks this somitic tendon progenitor population at its inception, and is continuously expressed through differentiation into the mature tendons.” The News Feature points out that my second sentence begins with a “brand new term”, *Scleraxis*, and quotes writing instructor Judith Swan as suggesting the following rewrite: “This somitic tendon progenitor population is marked at its inception by the gene *Scleraxis* (*Scx*)...” — reasoning that by first recapping the content of the previous sentence, the writer forms a “bridge” or transition that will better prepare readers to take in the new information.

I would first point out that while my second sentence indeed begins with a new term, it makes the transition, just five words later, back to the content of the previous sentence. More importantly, however, my decision to place *Scleraxis* at

the start of the second sentence, and to make it the subject of the sentence followed by an active verb, was to emphasize that it is only because of this marker and its continuous expression that identification of a somitic tendon progenitor population has become possible. Placing this information later in the sentence obscures the gene's importance to the research and cheats the reader of the excitement of discovery — of *Scleraxis* as well as the somitic tendon progenitor population it marks. In short, Swan's preferences, and mine, come down to a question of nuance and style.

While I agree that science writers should strive for clarity, they must be given the same latitude as writers in other disciplines to temper traditional rules of usage with individual stylistic choices that enhance what they want to communicate. Judging papers by methods such as LEX scores, which measure the ratio of everyday words to jargon, is, in my opinion, underestimating the capacity of science readers to creatively and flexibly handle the jargon, stylistic variations, and idiosyncrasies of interesting science writers. In fact, a bit of stylistic license and rule-bending, and a few new terms sprinkled here and there, might even keep the reader curious, challenged and awake. Clarity, yes; but clarity should not be synonymous with absence of style.

Ava Brent

Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA

Writing: the clear choice

Sir — It would be interesting to know the accessibility score for G. D. Gopen and J. A. Swan's paper “The science of scientific writing” (*Am. Sci.* **78**, 550–558; 1990), cited in your News Feature about clarity (*Nature* **423**, 376–378; 2003). I use this article in a course I teach. Although it makes some useful points, the students and I think it is overlong and full of unnecessarily complicated words and long-winded sentences.

Writing workshops are great but not widely available. As an alternative, I highly recommend the workbook *Essentials of Writing Biomedical Research Papers* by Mimi Zieger (McGraw-Hill, New York, 2000) — it works!

Sheila McCormick

Plant Gene Expression Center, USDA/ARS-UC-Berkeley, 800 Buchanan Street, Albany, California 94710, USA

correspondence

Contributions to Correspondence may be submitted to corres@nature.com. They should be no longer than 500 words, and ideally shorter.

Hydrogen lifts off — with a heavy load

The dream of clean, usable energy needs to reflect practical reality.

Paul M. Grant

For the energy-science community, the State of the Union address given by President George W. Bush this January contained a metaphorical hydrogen bomb. In a speech directed mainly at the woes of the US economy, the looming conflict with Iraq and the AIDS crisis in Africa, the president said: "Tonight I'm proposing \$1.2 billion in research funding so that America can lead the world in developing clean, hydrogen-powered automobiles. A single chemical reaction between hydrogen and oxygen generates energy, which can be used to power a car — producing only water, not exhaust fumes. With a new national commitment, our scientists and engineers will overcome obstacles to taking these cars from laboratory to showroom, so that the first car driven by a child born today could be powered by hydrogen, and pollution-free."

The White House later revealed the core programme to be a five-year, \$720-million initiative dubbed 'FreedomFUEL'. It is aimed at developing technology to support a national infrastructure for hydrogen production and delivery, and is directed principally towards the replacement of petroleum as individual vehicular fuel. The Bush administration deserves much praise for championing US determination to pursue aggressively the realization of what is surely the ultimate 'clean' fuel. We will soon see if Congress agrees — as the old adage goes: "the President proposes, but the Congress disposes".

The key technology for hydrogen-driven cars and trucks will probably be fuel cells. These have been under development for many years, and are beginning to show some signs of reaching maturity in terms of cost and performance. So let us concede that the day finally comes when your newborn drives an economical fuel-cell-powered vehicle with enough on-board hydrogen storage to travel 500 km without having to refill at a hydro-station too often. Where will the hydrogen come from? After all, this is fuel you can't mine or drill for.

The White House's *FreedomFUEL Fact Sheet* gives only one clue: "Hydrogen can be produced from abundant domestic resources including natural gas, coal, biomass, and even water." Well, these resources will have to be staggeringly abundant. For hydrogen to supplant petroleum completely for transport, its production would require an enormous outlay in capital plant and a significant



Clean getaway? George Bush has proposed funding to develop hydrogen-fuelled vehicles.

area of set-aside land. Even with a herculean effort on the model of the Apollo space programme, the hydrogen economy will arrive slowly, as it will require vast investment in infrastructure over a long period of time. This investment may not be driven by the market economy — it is more likely to evolve from a series of government energy-policy directives and legislative action that will mandate a blend of petroleum, methane, ethanol and hydrogen on the pathway to an eventual all-hydrogen transport supply. During this time there will certainly be a migration across several generations of technologies based on the internal-combustion engine, such as petroleum-electric hybrid vehicles.

Wasted opportunities

The United States uses almost 20 million barrels of petroleum every day, around 12 million of which power surface transport. A major portion of this is wasted, as internal-combustion engines have an efficiency of 20–30%. 'Only' three million barrels, therefore, are 'usefully' consumed. Because a fuel

cell is not a heat engine, and theoretically at least, escapes the bottleneck of the second law of thermodynamics governing internal-combustion motors, only enough homemade hydrogen is needed to offset the energy content of these three million.

Future fuel cells may be able to convert about 80% of the Gibbs free energy released by combining hydrogen with oxygen to make water into electrical energy (at present, this factor is around 50%). Also included in this should be the losses in both electricity conversion and electric-motor efficiency, around 20%, to 'shaft energy' to move the car. Thus the overall efficiency is 64%, much better than can be obtained from gasoline or diesel engines. So, we would need to generate around 230,000 tonnes of hydrogen daily — enough in liquid form to fill 2,200 space-shuttle booster rockets or, as a gas, to lift a total of 13,000 Hindenburg airships. Hopefully the thirst for this enormous quantity could be quenched by a factor of two or three by employing more efficient aerodynamic and drive-train designs in future hydrogen vehicles. But then folks would probably drive that much more.

Hydrogen is not a 'primal' energy source. Unlike fossil fuels or uranium, more energy is used to extract hydrogen from its source than is recovered in its end use. For simplicity, and to bypass issues of carbon and carbon dioxide sequestration, let us assume that the hydrogen is obtained by 'splitting' water with electricity — electrolysis. Although this isn't the cheapest industrial approach to 'make'

AP PHOTO/CH PEDRONCELLI

REUTERS/LARRY DOWNING

hydrogen, it illustrates the enormous production scale involved — about 400 gigawatts of continuously available electric power generation have to be added to the grid, nearly doubling the present US national average power capacity. The number of new power plants that would need to be built — based on presently available technologies — to meet this demand is roughly 800 natural-gas-fired combined-cycle units generating 500-megawatts, or 500 800-megawatt coal-fired units, 200 Hoover Dams (two gigawatts each), or 100 French-type nuclear clusters (four reactors, about one gigawatt each).

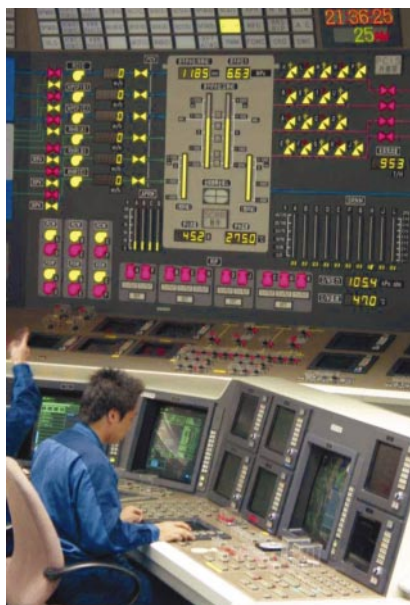
The average capital cost of building an electric power plant is \$1,000 per kilowatt (with considerable variance), which would mean new investment of at least \$400 billion (one-twentieth of US gross domestic product). This does not include the storage and delivery costs that would be incurred for a complete transformation to a surface-transport system running on hydrogen instead of petroleum. A daunting prospect, but not impossible. To get the daily hydrogen ration of 230,000 tonnes, just over two million tonnes of water is required. Even this vast amount of water expelled as 'exhaust' will be recycled to the environment in several days, unlike carbon dioxide.

A popular justification for moving from petroleum to hydrogen, one that is often invoked by US political pundits, is 'freedom from dependence on Middle East oil'. Yet according to the latest figures published by the US Energy Information Agency, only 14% of US oil imports come from Gulf states — Europe (30%) and Japan (75%) are much more reliant on these resources.

For me, the most compelling rationale for the hydrogen economy is the potential to drastically reduce carbon emissions. If we can get energy without oxidizing carbon, I believe it's a good idea to do it. According to studies by the Electric Power Research Institute in Palo Alto, California, about 97% of the hydrogen produced in the United States is derived from the thermocatalytic 'splitting' of natural gas or refinery gases, or 'coal gasification' — the reaction of water (steam) with carbon to yield hydrogen and carbon monoxide. The heat to effect both of these processes is based on fossil combustion. To derive hydrogen electrolytically for transport would therefore yield carbon emissions likely to offset most of the benefits of saving nine million barrels of petroleum per day.

Reality bites

What about using renewable energy resources for this job? Sweden's Norsk Hydro currently builds the largest electrolyzers, and British Colombia-based firm BC Hydro recently completed a pilot project that is now making hydrogen for use in fuel-cell-powered buses in Vancouver. But I doubt whether we will see even one major new North American



Nuclear know-how: the Kashiwazaki-Kariwa plant is a highly efficient new generation facility.

hydroplant built in the near (or far) future expressly to generate hydrogen, let alone 200 Hoover Dam equivalents.

Wind and photovoltaic solar technologies are expected to approach power densities of about 10 and 100 watts per square metre, respectively, when the wind is blowing hardest and the Sun shining the brightest. Imposing a realistic annual duty factor of 20% on Sun and 30% on wind, the Earth surface area required to produce 400 gigawatts would be 130,000 km² for wind and 20,000 km² for solar technologies. The former is about the size of New York State, the latter about half the size of Denmark.

Another option, extracting energy from biomass through combustion, is attractive in principle, as the carbon emitted is recycled through future plant growth. About 15% of the land area in the United States is under cultivation, mostly for food. Some estimates put the annual biomass energy equivalent of these crops at roughly 23 million gigawatt-hours, or about 2,600 gigawatts of continuous power production¹. To add another 400 gigawatts of electrical power to produce hydrogen for transport would require another 3% to undergo cultivation, an area about the size of Nevada.

Whenever you do something 'massive' to ecology — such as pumping gigatonnes of carbon dioxide into the atmosphere, or proposing million-acre wind or solar farms — the environmental consequences are uncertain at best and disastrous at worst. An example that relates to carbon recovery or sequestration is the experiment that used iron to fertilize the oceans near Antarctica in an attempt to capture carbon². This resulted in huge algal blooms followed by emission of methyl bromide, an ozone poison.

If there is a way to make both electricity and hydrogen without releasing carbon dioxide, it would probably be wise to use it. In fact, it does exist and is already in use. I believe that its science is sound and solved, its technology mature and safe. It is called nuclear fission power. Its fuel costs are fixed and supply is assured for at least 500–1,000 years, by which time there may be neutron-free, 'clean' fusion — maybe. Readers accustomed to the unscientific belief that nuclear power is inherently dangerous and forever unsafe should examine the facts, starting with the article "The Need for Nuclear Power"³.

Sensible solutions?

On the west coast of Japan, at Kashiwazaki Kariwa, the Tokyo Electric Power Corporation has spent 20 years building perhaps the world's most modern nuclear power complex. Its reactor units generate eight gigawatts of electric power, available 90% of the time, on a site that occupies slightly less than 4 km². The site includes a visitor centre, machine shops and on-site 'spent' fuel storage. More than a quarter of the plant area is green space. Sand dunes and native growth are left undisturbed. Yet the power density is still an amazing 1,800 watts per square metre. Thus, the 400 gigawatts of electricity needed to extract hydrogen from water to power US surface transport could be cumulatively generated on land occupying only 233 km².

Any form of highly concentrated potential energy, from dynamite to deuterium, will provide many opportunities for humans to misbehave and make mischief. Nonetheless, I believe that a resurgence of nuclear power is necessary for the continuing industrialization of world society with minimal environmental impact and eco-invasion, one in which hydrogen will supplant fossil sources. FreedomCAR and FreedomFUEL cannot be had without FreedomNUKES — and FreedomNUKES cannot be had in a world that continues to permit the unrestricted proliferation of nuclear weapons. International laws and institutions must be established to control and vigorously enforce the use of actinide materials for peaceful purposes from minehead, through recovery and breeding, to eventual disposal, and to prevent diversion to rogue weapons programmes. Only then can the vision that the parents of the atomic age foresaw and desired be realized — a world where 'atoms for peace' would prevail, creating a clean energy source independent of any geographically accidental richness of fossil reserves. ■

Paul Grant is a science fellow at the Electric Power Research Institute, PO Box 10412, 3412 Hillview Avenue, Palo Alto California 94303, USA.

1. Ristinen, R. A. & Kraushaar, J. J. *Energy and the Environment* (Wiley, New York, 1999).

2. Boyd, P. W. *et al. Nature* **407**, 695–702 (2000).

3. Rhodes, R. & Beller, D. *Foreign Affairs* **79**, 30 (2000).

A dish to die for

Apes are still killed for their meat, but awareness of their plight is growing.

Eating Apes

by Dale Petersen

University of California Press: 2003. 333 pp.

\$24.95, £16.95

Guy Cowlshaw

The trade in wild meat is now perceived as one of the most important threats to global biodiversity. Nowhere is this more likely to be true than in tropical Africa, where the unsustainable hunting of wild animals has led to a 'bushmeat crisis' that threatens myriad forest species with extinction. Those most at risk are the great apes: gorillas, chimpanzees and bonobos (pygmy chimpanzees). Already living in small populations, the large body size of the apes makes them particularly attractive to hunters, while their slow reproductive rates make them unable to sustain even a modest offtake. Their biological characteristics are, quite literally, a recipe for disaster.

Despite early warnings in the 1970s and 1980s, the bushmeat trade only captured the attention of the wider international community in the mid-1990s. Part of the credit for this change in awareness is due to the photographer Karl Ammann, and his efforts to publicize the bushmeat trade through popular articles, television and an animal-welfare campaign. But it is Ammann's photographs that have been most influential. His dramatic images of butchered apes brought the bushmeat trade out of the African rainforest and presented it for scrutiny to the rest of the world.

Dale Petersen's *Eating Apes*, written with the assistance of Ammann (who also supplies the afterword), charts the history of Ammann's first exposure to the bushmeat trade and his role in bringing the trade to the attention of the Western public. As such, it is a fascinating account of some of the key people and events that have contributed to the development of one of today's most significant environmental movements. The underlying theme of Petersen's book, however, is not biographical or historical, but ethical: "Is it right, in the deepest moral sense, for one conscious being to eat another?"

The book begins with an exploration of the rich intellectual and emotional lives of our closest relatives. From this starting point, the narrative unfolds with a mixture of biography and science. Episodes in the life of Ammann and his associates are punctuated with numerous facts and figures about the bushmeat trade: accounts of how apes are hunted, the nutritional value of meat, the principles of sustainability, popular recipes for cooking gorilla, and perhaps most notably, the possi-



Hand to mouth: a gorilla hand in a Congolese restaurant throws the bushmeat trade into sharp relief.

bility that HIV/AIDS originated from simian immunodeficiency viruses that leapt from nonhuman primates to people through the hunting and handling of bushmeat.

The scale and complexity of the bushmeat crisis make it peculiarly difficult to identify a single underlying cause. Current thinking suggests a variety of factors, including human population growth, forest loss, greater accessibility to remote forest areas, the adoption of new hunting technologies (such as firearms and wire snares) and an increasing commercialization of the trade. Petersen focuses particular attention on logging operations, which can facilitate and promote the local bushmeat trade in all these ways. To tackle this problem, a number of conservation agencies have recently begun to work in collaboration with logging companies. These initiatives have met with mixed results. Through a detailed dissection of one such partnership, Petersen highlights his concerns about these schemes, namely transparency (a lack of independent assessment), equitability (a disproportional financial burden on the conservation partner) and exploitation (a public-relations victory for the logging industry). Petersen's conclusion is not that the collaborative approach should be abandoned, but rather that improved implementation is desirable. Indeed, one might imagine these difficulties as the inevitable growing pains of any new initiative to combat such a complex problem.

By following the progress of Ammann and his colleagues over the years, and by recounting both their victories and misadventures, Petersen provides a personal view of the bushmeat trade that is illuminating and eminently readable. The drawback of this approach is that it only represents one view of the current crisis. Petersen and Ammann are driven largely by concerns of welfare and the morality of hunting and eating apes. The conservation perspective, concerned with issues of wider biodiversity and ecosystem viability, is given considerably less coverage. The humanitarian perspective, concerned with the loss of income and food security to the bushmeat-dependent rural poor, is barely mentioned. Petersen's book should therefore not be considered as a definitive treatment of the bushmeat phenomenon.

Eating Apes is a strange stew — an emotive account of the bushmeat trade, and the threat it poses to the African apes, that is both popular and polemical. As such, it may not be to everybody's taste. Nevertheless, it is an absorbing mixture of biography, biology, anthropology, politics, economics and ethics. It is also thoroughly researched (extensive notes are provided for each chapter). Fittingly, it is illustrated by a selection of Ammann's most famous photographs. ■

Guy Cowlshaw is at the Institute of Zoology, Zoological Society of London, Regents Park, London NW1 4RY, UK.

K. AMMANN

Design of paradise

A Magic Web

by Egbert Giles Leigh Jr &
Christian Ziegler

Oxford University Press: 2003. 292 pp. \$40

Lawren Sack and Dina Dechmann

Tropical rainforests are the settings for innumerable interactions between millions of organisms, a centrepiece of evolution. A *Magic Web* showcases the complexity of these interactions, using the rainforest that has been studied more than any other in the world, that of Barro Colorado Island (BCI), a 15-km² island in the Panama Canal.

Its author, Egbert Leigh, has worked on BCI since the 1960s, and is especially concerned about what creates and supports tropical diversity. Here he summarizes data collected by hundreds of investigators, and argues that the finely tuned web of interactions between organisms is an important basis for the species diversity and complexity of tropical rainforest. Leigh points to many differences among species that are driven by direct competition. He argues convincingly for an equal importance of interactions between animals and plants. Plants diversify in their ways of escaping herbivores and enticing animals to pollinate their flowers and disperse their seeds. Animals diversify in the way they subsist on plants and avoid being eaten by other animals. It is a dynamic web; changes in climate can alter plant processes, with knock-on effects on the food available to animals. Loss of species, often induced by humans, unravels the web.

Leigh's wide-ranging vision truly comes alive when it meshes with Christian Ziegler's striking images. Ziegler has a degree in tropical biology and spent 15 months on the island photographing for this book. His collection of the typical as well as exceptional sights of BCI reveals a great intimacy with the forest. Ziegler does not simply illustrate the science, but rather illuminates each topic discussed, leading to a strong sense of the interactions everywhere in the jungle. Highlights include the sections on forest light gaps and plant diversity, on ant colonies, and on animal mimicry and camouflage.

The book will appeal to a broad audience. As a coffee-table book, it is enjoyable for its photographs and captions alone. The main text provides a useful introduction to tropical rainforest, surveying a wide range of ecological concepts — although, of necessity, briefly and in simple terms. These concepts are grounded with entertaining metaphors from human society. There are numerous comparisons with other ecosystems, and 'back of envelope' calculations. Ultimately, the book will enrich any scientist's view of biology. The tropical biologist will crave returning to the field.



Forest food: a broad-billed motmot arrives home bearing an unfortunate beetle for its chicks' dinner.

Of course, even such a detailed treatment can be only a rough sketch of the intricate network of tropical-rainforest life. Leigh jokes that humans are not distinguished for their omniscience, and Ziegler reminds us how elusive are the visions of nature that he depicts. To fathom the web in all its detail, we will need to step up collaborations between scientists with diverse expertise and approaches. BCI, with its terrific facilities, has helped to

build many bridges. Leigh and Ziegler's colourful synthesis should inspire many more to the task.

Lawren Sack is in the Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Avenue, Cambridge, Massachusetts 02138, USA.

Dina Dechmann is in the Department of Animal Behaviour, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland.

Sex on the brain

Essential Difference: Men, Women and the Extreme Male Brain/

Essential Difference: The Truth

About the Male and Female Brain

by Simon Baron-Cohen

Allen Lane: 2003. 256 pp. £16.99/

Perseus: 2003. 256 pp. \$26

Joyce F. Benenson

How male and female brains adapted differently to the environment in which humans evolved is an intriguing question. The traditional response from evolutionary scientists is that females' greater childcare responsibilities and males' greater involvement in intra-group and inter-group competition for resources and mates produced adaptations specific to these domains.

Baron-Cohen boldly suggests broader adaptations. Forming the foundation of his theory are his rich descriptions of Asperger's syndrome, a condition that affects more males than females by a ratio of up to 10:1. Individuals with Asperger's syndrome,

which is a mild form of autism, experience difficulties forming social relationships and communicating with others, and exhibit rigid and repetitive behaviours. Baron-Cohen interprets the symptoms of Asperger's syndrome as indicating an overabundance of systemizing and a lack of empathizing skills. On the basis of his clinical work with individuals with this condition, as well as his empirical research with unaffected children and adults, he infers that sex differences between typical males and females go beyond differential investment in children versus mating or hunting or warfare. Instead, he proposes that the male brain is programmed to systemize and the female brain to empathize, and that Asperger's syndrome represents the extreme male brain.

The book is written for a general audience, and it is easy and enjoyable to read. Case studies are intertwined with more systematic analyses. However, it is not the place to look for a comprehensive overview of empirical research in the field of either sex differences in personality or Asperger's syndrome. Because of the novelty of his ideas, Baron-Cohen defines his constructs on the basis of

scales he has created and generously included in their complete forms in the appendices. Consequently, each reader can determine his or her own systemizing quotient, empathy quotient and autism quotient, as well as estimate those for one's family members, friends and less-liked associates.

The idea that males are more interested in systemizing than females merits serious consideration. Is there a common underpinning to such diverse pursuits as generating taxonomies of plants and animals, throwing balls or darts, skiing, figuring sports averages, creating maps and charts, designing buildings, repairing automobiles and houses, generating art, concocting culinary dishes, playing instruments, and studying mathematics, physics and engineering? Could it be that males engage in these behaviours more than females because of some heightened attraction to hypothesis testing and classificatory schemes? The question delves into the very nature of human intelligence itself. Because males typically ventured farther from their homes than females did in the environment in which humans evolved, those males who were able to systemize with greater precision might have gained an evolutionary advantage. Females who were more engaged than males in interaction with immediate family members in familiar environments might not have benefited so strongly from this type of skill. It is unquestionably a novel and fascinating idea that seems likely to generate a rich empirical body of literature as its properties are tested.

The second part of the theory — that females are more empathic than males — is more problematic. Baron-Cohen believes that females are better than males at caring about and communicating with others — that is, at reading others' emotional cues and responding appropriately. To support this part of his theory, Baron-Cohen cites studies showing that, beginning in childhood, compared with males, females make more eye contact, converse more politely, are less overtly competitive, engage in less physical aggression, and as adults, commit fewer homicides and care more for children and other helpless individuals.

Other measures, however, show that males are highly socially skilled. Developmental research shows that, beginning in middle childhood, males are more likely than females to cooperate in same-sex groups and to include newcomers. Furthermore, despite the lower verbal intimacy between males than among females, males' friendships endure for longer. Anthropological studies show that as adults, males are more likely than females to forge strong political, military, financial, educational and religious alliances. These relationships require both a high level of investment in others as well as excellent communication skills, including the ability to decipher subtle

Science in culture

Optical alchemy

Bridget Riley's constructions create ripples in the flat world of geometry.

Colin Martin

For 40 years, Bridget Riley, pioneer of 'op art' (optical art) has used simple geometric shapes to make complex, sometimes visually unsettling paintings. As viewers' eyes move across a typical canvas — its lines, spots, circles and squares — they perceive movement. Adjacent areas advance and recede.

Rather than using optical theory, Riley creates these visual effects intuitively. She refines her preparatory drawings until she achieves her desired result, which is to evoke emotion through formal abstraction. When satisfied, she entrusts the painting of finished canvases to studio assistants.

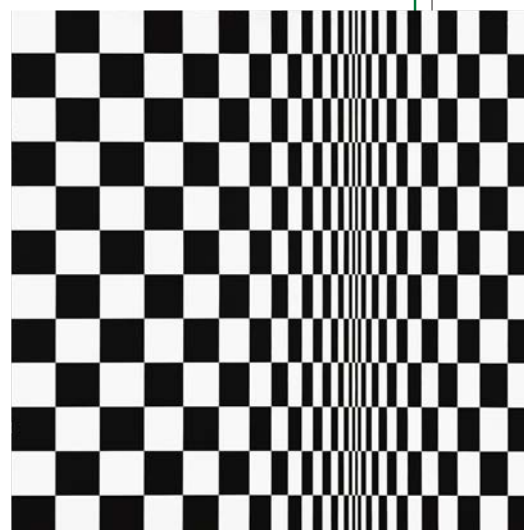
Her account of altering a single variable, the width of a square — in *Movement in Squares* (1961, right) — is reminiscent of experimental science. "I took a simple element [and] decided to change one of its characteristics, leaving the others intact. The result is a dramatic increase in tension. As the square comes under extreme pressure, a tremendous visual vibration is set up. In this critical area the contrast between blacks and whites is intensified and dazzling."

Movement between repose to disturbance is a constant feature in Riley's early, black-and-white work. Subsequently, she introduced colours — using straight and then curved lines — producing the chromatic fields that result from interactions between colours. Throughout her career, Riley has returned to her earlier themes, as composers do in

writing music. The beautiful rhythm that unfolds across 2003 canvas *Evoë 3* (bottom), in which she sets curved shapes in pink and magenta against a blue and green ground, recalls the lyrical movement of Matisse's dancers.

Colin Martin is at 32 Woodstock Road, London W4 1UF, UK.

A retrospective exhibition of Bridget Riley's work can be seen at Tate Britain until 28 September 2003.



emotional signals and to respond appropriately. The fact that males are less verbally polite, discuss fewer personally intimate topics, and engage in more overt competition and aggression to achieve personal goals does not preclude them from forming strong bonds. On the contrary, males respect and desire to associate with other males who are powerful. My own research indicates that females and males excel at different types of social skill.

The link between Asperger's syndrome and the male brain, then, may work well for systemizing skills. However, the lack of interest in social relationships demonstrated by these individuals is inconsistent with the

high level of social skills displayed by many males. Perhaps Baron-Cohen will find that individuals with Asperger's syndrome engage in the types of social skill displayed more by males than by females. If not, individuals with Asperger's syndrome might represent only one form of the extreme male brain. Another form would include males who score highly on systemizing skills and who are also highly sociable. Regardless of the ultimate answer, this book inspires the reader to reconsider traditional assumptions about the skills of each sex. ■

Joyce F. Benenson is in the Department of Psychology, University of Plymouth, Drake Circus, Plymouth, Devon PL4 8AA, UK.

Tiny teamwork

E. Peter Greenberg

Until recently, bacteria were considered to be self-contained and self-sufficient individuals. These unicellular organisms were thought to lack the sophistication of plants and animals to organize into multicellular groups. We also assumed that they lacked the ability to communicate, a crucial function for organizing group activities. Our view has changed. Bacteria can organize into groups, they can communicate, and these abilities are important factors in the development of many diseases. Organized groups of bacteria in the form of biofilms often cause persistent infections, such as those of the middle ear, urinary tract, bone, heart valves and implanted medical devices. Because we have not considered the problem of group biology in bacteria until recently, we lack good therapeutic strategies to treat biofilm infections.

The first hints that bacteria might produce species-specific signals for cell-to-cell communication emerged more than 30 years ago from pioneering work in the laboratories of J. W. Hastings and Alex Tomasz. When critical environmental concentrations of then unidentified signalling molecules were reached, they triggered the expression of specific genes, controlling light production in a marine luminescent bacterium, and genetic competence in the pneumococcus. After a grudging acceptance of these examples, communication was viewed as a curiosity, unique to a few special bacteria. But we now know that many bacteria use cell-to-cell communication to control gene expression, a process that has become known as quorum sensing. At a high population density, there is enough signal to instruct all of the individuals to do something they won't do at lower population densities — a form of peer pressure.

Quorum sensing in *Pseudomonas aeruginosa* has attracted considerable attention because this pathogen causes many hard-to-treat infections, including chronic biofilm

infections in the lungs of people with cystic fibrosis. In this bacterium, quorum sensing is a key virulence determinant — mutants that cannot communicate are seriously impaired in their ability to cause infections in lab models. *P. aeruginosa* uses quorum sensing to control hundreds of genes, many of which code for the production of secreted virulence factors that are toxic in one way or another to the host.

Both the signalling molecules and the genes expressed differ between species. For example, the pneumococcus and several other bacteria use a peptide as their signal, whereas *P. aeruginosa* and the luminescent bacterium rely on acylated homoserine lactones.

The realization that many bacteria can communicate, and that there are different systems for communication, has led to what seems like a gold rush among microbiologists. But it is important to distinguish true communication from more general responses. In humans, there is an obvious distinction between communication and a startle response to a loud noise, but with bacteria the situation looks less clear-cut. There is evidence that some bacteria can measure a signal produced by most species and respond to the general bacterial population density. Is this signalling and response, but not communication? How could this have evolved?

In parallel with the new field of quorum sensing, great strides were also made in understanding surface-attached groups of bacteria — biofilms. Advances in microscope technology for imaging living biofilms allowed us to observe that bacteria growing on surfaces can develop into structured but heterogeneous groups. Even within groups of a single species, individuals in different locations have different activities. For example, when grown on a surface, *P. aeruginosa* can form a forest of mushroom-like structures and not just a uniform blanket of growth. Using fluorescent stains and micro-electrodes, investigators showed that different regions within the structures had different metabolic activities.

Finally, in 1996, the fields of quorum sensing and biofilms converged when we found that *P. aeruginosa* mutants with quorum-sensing defects failed to form the mushroom-like structures. A connection between quorum sensing and biofilm development has since been found in several other bacterial species. Quorum sensing seems to promote biofilm development in *P. aeruginosa* and some other pathogens, such as *Burkholderia cepacia*, whereas in other bacteria (for example, *Vibrio cholerae* and the photosynthetic bacterium, *Rhodobacter sphaeroides*) it might have a role in the dispersal of individual cells from a biofilm. Although the quorum-sens-

Bacterial communication

Recognition that bacterial cells can communicate and organize into groups has led to new ways of thinking about chronic infections.

ing-biofilm connection is evident in several bacteria, it is important not to overstate its significance. Bacterial species that are not known to communicate can form biofilms, and even *P. aeruginosa* can form biofilms, albeit abnormal ones, without quorum sensing. Perhaps the true significance of the quorum-sensing-biofilm connection is that it points to a genetic component of the development of structured biofilms.

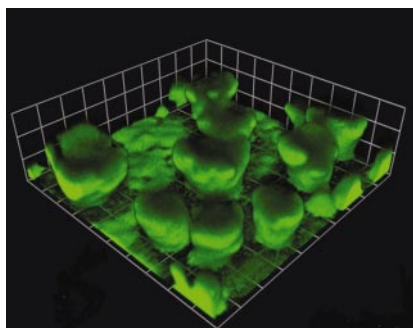
With recent progress in understanding bacterial communication and communities has come an appreciation that biofilms are an important problem in medicine. Bacteria residing in biofilms can resist host defences, and they show a tolerance to antibiotics at concentrations that would kill them outside the biofilm. This helps to explain why biofilms can cause devastating chronic infections like the *P. aeruginosa* lung infections that plague and ultimately kill people with cystic fibrosis. A better basic understanding of bacterial group activities should lead to the development of new therapies for the treatment of lung infections in cystic fibrosis, and of other biofilm-based diseases.

With all of the promise and excitement of this new area in microbiology, there are still basic knowledge gaps about the prevalence of communication and community behaviour in bacteria, and about the evolution and diversity of these phenomena. Microbiologists need to learn from population biologists and ecologists who have been thinking of and working on communities and communication for decades. Conversely, bacteria may serve as convenient models for population biologists and ecologists to test hypotheses about the rules governing the evolution of cooperative and group activities. ■

E. Peter Greenberg is at the Carver College of Medicine and in the Department of Microbiology, University of Iowa, Iowa City, Iowa 52242, USA.

FURTHER READING

- Davies, D. G. *et al.* *Science* **280**, 295–298 (1998).
- Costerton, J. W., Stewart, P. S. & Greenberg, E. P. *Science* **284**, 1318–1322 (1999).
- Fuqua, C. & Greenberg, E. P. *Nature Rev. Mol. Cell Biol.* **3**, 685–695 (2002).
- Whitehead, N. A., Barnard, A. M., Slater, H., Simpson, N. J. & Salmond, G. P. *FEMS Microbiol. Rev.* **25**, 365–404 (2001).



Culture club: *Pseudomonas aeruginosa* can gang together to form mushroom-shaped structures.

Catalyst hunt accelerates

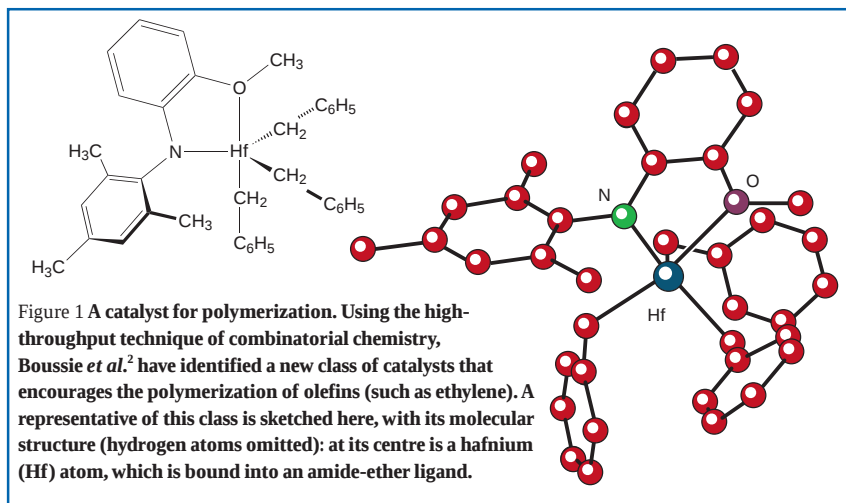
Virgil Percec

High-throughput methods of synthesis and analysis are streamlining the search for new catalysts, reducing the timescale from years to days. A new class of polymerization catalysts is the latest discovery.

The discovery of new catalysts, or the means of improving existing ones, is the cornerstone of the highly competitive chemical and pharmaceutical industry. Yet, traditionally this process is largely an empirical, time-intensive matter of trial and error. For example, in the commodity chemicals business the discovery and implementation of new catalytic processes for the production of polymers such as polyethylene can take decades¹. High-throughput experimentation, or 'combinatorial chemistry', however, aims to increase the rate at which a researcher can design, perform and evaluate experiments over conventional one-at-a-time methods. Hundreds, even thousands, of experiments can be screened rapidly, and often simultaneously. To accommodate such experimental throughput, the experimental format is often miniaturized, perhaps with the aid of robotics. Rapid analysis of a large number of experiments generating extremely small amounts of experimental products is then the challenge to be faced.

The collaboration of Symyx Technologies and Dow Chemical, reported by Boussie *et al.*² in the *Journal of the American Chemical Society*, has come up with a novel and effective high-throughput technology for finding catalysts for the production of polyolefins with targeted properties. Polyolefins are polymers of simple olefins such as ethylene, and α -olefins such as propylene and 1-octene. The polyolefins industry is by far the largest polymer business in the world, and the drive to improve the properties of these materials through careful control of their structure has fuelled the search for appropriate organometallic catalysts. These catalysts have a transition-metal centre, stabilized by an organic molecule (a ligand), but their performance is profoundly and unpredictably influenced by the nature of the metal and ligand, as well as by experimental conditions. Boussie *et al.*² have demonstrated the remarkable power of their high-throughput approach by discovering a new class of hafnium-based catalysts that are capable of producing high-molecular-mass copolymers of ethylene with 1-octene at extremely high temperatures.

Dow currently manufactures more than half a billion kilograms of such polymers annually, in a solution process for which the production costs are significantly lower at



higher temperatures. Hence there is an incentive to find catalysts with high-temperature performance. But to compare the efficacy of different catalysts, each catalyst must be tested under conditions that produce copolymers with the same ratio of ethylene to α -olefin. This requires numerous experiments because of the inherent variation in the catalyst's propensity to incorporate each olefin into the polymer chain.

To overcome the limitations of this traditional approach, Symyx has developed a staged screening approach in which each stage becomes progressively more discriminating than the previous one. The initial (primary) screening involves evaluating large arrays of catalyst candidates for their ability to homopolymerize an α -olefin such as 1-octene. Because it is typically more difficult to polymerize an α -olefin than ethylene, this primary screening identifies catalyst candidates that also have the potential to copolymerize ethylene and α -olefin monomers, and provides an inferential means of assessing what molecular mass of polymer can be built up. 1-Octene is a liquid and so can be handled easily, and it produces a soluble polymer product after polymerization, thereby facilitating downstream characterization of the product.

For the primary screening, 384 experiments consisting of combinations of two transition metals (zirconium and hafnium) with 24 ligands and eight unique activation conditions were performed at room temperature in a matter of hours. Screening such large numbers of experiments required

an ability to perform polymerizations on a small scale and to characterize small amounts of polymer products. The technology developed to do this consists of arrays of 1-millilitre glass vials containing reaction volumes of 250 microlitres and tiny amounts of catalyst (about 400 nanomoles), and the use of a variety of proprietary techniques to assess their molecular mass and composition³⁻⁵.

The most promising candidates identified during the primary screen were then selected for the next stage of (secondary) screening on the basis of the yield and molecular mass of poly(1-octene) produced. Of the 384 experimental combinations screened, 10 displayed activities worth secondary screening, initially in the presence of both ethylene and 1-octene and at 130 °C to assess their potential as high-temperature copolymerization catalysts. For the secondary screening, a specially designed array of miniaturized reactors was constructed. The 'parallel polymerization reactor'⁶ houses an array of 48 small reaction vessels and has many of the features of a traditional batch reactor including mechanical stirring, temperature control and monitoring, as well as the ability to measure the uptake of gases such as ethylene in a relatively small sample volume. Each well within the reactor is designed to accommodate a reaction volume of about 5 millilitres — between 100 and 1,000 times smaller than a standard laboratory autoclave (high-pressure chamber).

Using this technology, Boussie and colleagues² found that, under a particular set of

polymerization conditions, the combination of hafnium with an amide-ether-based ligand gives a catalyst that meets the requisite high-temperature copolymerization targets (Fig. 1). Members of this class of catalyst produced a good yield of the ethylene–1-octene copolymer and met molecular-mass targets. To explore the potential of this type of ligand, and also to optimize the catalyst structure as well as experimental conditions, a further high-throughput study of 96 amide-ether ligands was performed. Boussie *et al.* found that the catalyst performance varied tremendously and in an unpredictable fashion, thus endorsing the benefits of the high-throughput approach. Samples of this new class of polymerization catalyst were also then sent to Dow for further validation using its full-scale batch reactors. The results were corroborated; in fact, the Dow analysis found that the new hafnium catalysts rival some of the company's own commercially relevant catalysts.

Boussie *et al.*² have shown that high-throughput strategies can indeed accelerate the discovery of new catalysts⁷. The materials community can now exploit the rapid synthetic strategies that are already familiar to

pharmaceutical chemists⁸. But, to facilitate the rapid discovery of complex molecular and supramolecular materials, even more sophisticated strategies of rapid synthesis and structural analysis will be needed^{9–11}. The challenge of discovering in a matter of days what used to take years will not stop here.

Virgil Percec is in the Roy & Diana Vagelos

Laboratories, Department of Chemistry, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6323, USA.

e-mail: percec@sas.upenn.edu

1. Ewen, J. A. in *Metallocene-Based Polyolefins* (eds Scheirs, J. & Kaminsky, W.) 3–31 (Wiley, New York, 2000).
2. Boussie, T. R. *et al.* *J. Am. Chem. Soc.* **125**, 4306–4317 (2003).
3. Nielson, R. B., Kuebler, S. C., Bennett, J., Safir, A. & Petro, M. US Patent 6,175,409 (2001).
4. Boussie, T. R. & Devenney, M. European Patent Application 1-160-262-A1 (2001).
5. Komon, Z. J. A. *et al.* *J. Am. Chem. Soc.* **124**, 15280–15285 (2002).
6. Turner, H. W., Dales, G. C., Van Erden, L. & van Beek, J. A. M. US Patent 6,306,658 (2000).
7. Schultz, P. G. & Xiang, X.-D. *Curr. Opin. Solid State Mater. Sci.* **3**, 153–158 (1998).
8. Kohli, R. M., Walsh, C. T. & Burkart, M. D. *Nature* **418**, 658–661 (2002).
9. Percec, V., Cho, W.-D., Ungar, G. & Yeardley, D. J. P. *J. Am. Chem. Soc.* **123**, 1302–1315 (2001).
10. Percec, V. *et al.* *Nature* **419**, 384–387 (2002).
11. Lehn, J.-M. *Proc. Natl Acad. Sci. USA* **99**, 4763–4768 (2002).

but such infection clearly occurs *in vivo*^{5,6}.

Into this complex mix of studies comes the paper by Swingler and colleagues¹. They have carried out an exciting series of experiments that provide an indication of the complexity involved in developing a permissive viral reservoir in relatively inactive T cells. As depicted in Fig. 1, the authors show that HIV-1 infection of macrophages leads to the production of at least two soluble factors, sCD23 and sICAM-1. These factors then stimulate the production of accessory surface molecules on another component of the immune system, B cells, when the B cells are in close proximity to the HIV-1-infected macrophages — that is, when the two cell types are occupying the same microenvironment. When, in turn, these B cells come into contact with non-cell-cycling T cells, they allow the T cells to become permissive for HIV-1 infection. These events can occur without the T cells being fully activated and in the absence of cell-cycle activity.

What is especially interesting about these studies¹ is the demonstration that the HIV-1 accessory protein Nef is involved. Nef is a key viral protein in disease induction *in vivo* by both HIV-1 and SIV, acting in the pathways that induce changes in the activation state of an infected cell^{7,8}. When expressed from HIV-1, or even alone from another non-HIV-1 viral vector, Nef leads to increased production of sICAM-1 and sCD23 molecules from infected macrophages. This is similar to the pathway of induction of selected macrophage-derived factors from the natural ligand CD40 found on uninfected macrophages⁹. In some cases, Nef may even be produced directly in 'pre-integration latently infected' T cells^{10,11}. So Nef seems to engage in molecular piracy, taking over an existing cellular pathway and allowing HIV-1 to replicate in T cells in diverse activation states.

This would be complicated enough, but it is only part of the story. It seems that sCD23 first has to alter B cells before they in turn can make non-cell-cycling T cells into fertile

HIV

Cross-talk and viral reservoirs

Roger J. Pomerantz

Even after intensive antiviral treatment, HIV-1 can lurk in reservoirs in the human body and then reappear. The complex pathways by which the virus creates such reservoirs are only now becoming clear.

Human immunodeficiency virus type 1 (HIV-1) infects many cells in the human host, but there are two major targets — macrophages and CD4⁺ T cells. Over the years, the list of complicated ways in which the virus is known to behave has grown. In their paper on page 213 of this issue, Swingler *et al.*¹ add to the list. They show how, through stimulating interaction between various host cells, HIV-1 sets up new potential viral reservoirs in 'resting' T cells.

At one time, the dogma was that HIV-1 could infect only actively replicating T cells that are going through the cell cycle. These are cells that have left the resting stage on encountering antigen of an invader, and in response have become activated and have started to reproduce themselves. But this dogma has been shown to be overly simplified. For example, low levels of the intercellular signalling molecules known as cytokines are enough to produce HIV-1 infection of resting human T cells². The evidence of this induction of 'viral permissivity' in resting cells has meant that cytokines such as interleukin-7 have received much attention of late³. Complementary *in vivo* data

have come from studies with both HIV-1 and simian immunodeficiency virus (SIV). From these it seems that T cells that have not been fully activated can replicate SIV and HIV-1, and act as low-level producers of virus and as residual viral reservoirs in infected monkeys and humans, respectively⁴. How exactly HIV-1 infects resting cells has been controversial,

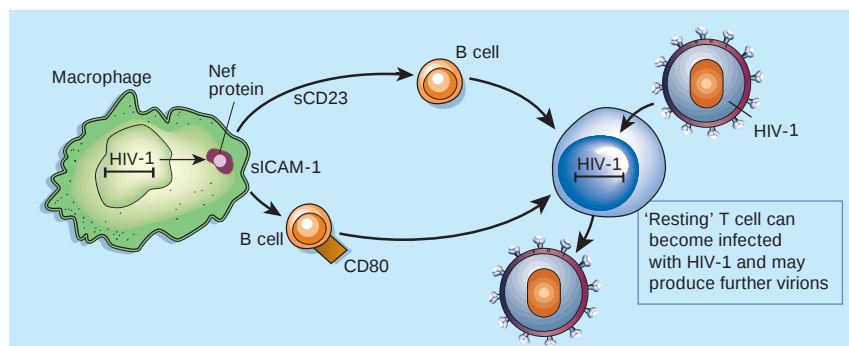


Figure 1 Cell-to-cell cross-talk, and the induction of HIV-1 infection in resting CD4⁺ T cells. The sequence of events described by Swingler *et al.*¹ begins with the production of the viral accessory protein, Nef, in HIV-1-infected macrophages. Nef prompts the production of soluble factors (sICAM-1 and sCD23), which stimulate B cells to become the intermediates in making resting, non-cell-cycling T cells become 'permissive' for viral infection. Finally, viral replication and production can also be induced in certain non-cell-cycling T cells.

fields for HIV-1 infection and replication. Yet another signal — CD80, induced by sICAM-1 — has to be expressed on the surface of B cells to allow productive replication of HIV-1 in the target T cells, rather than just allowing infection. Otherwise, the T cell becomes infected but has a provirus that may act as a 'Trojan Horse of non-production'¹², allowing viral expression and release to take place only after a second signal, such as CD80, is applied from a further activated subset of B cells. In previous work, the Swingler group¹³ had demonstrated that HIV-1 Nef produced from macrophages may, through chemotaxis, also actually attract T cells into physical proximity with the infected macrophage.

If this sounds complex, it is — even for experts in the study of retroviruses. But no one said that understanding HIV-1's reservoirs would be simple. The data presented by Swingler *et al.* are unique in that they demonstrate that viral reservoirs during highly active antiretroviral therapy in HIV-1-infected patients may be far more intricate than consisting simply of low-level replication in activated T cells and latency in fully resting T cells¹². These new data indicate ways in which productive replication from certain viral reservoirs can be upregulated, and also may explain how some non-cell-cycling T cells can become a latent but inducible viral reservoir *in vivo*. Overall, they show that there is an entire spectrum of interactions between the virus and host cells,

which together produce the general pattern of HIV-1 reservoirs and residual disease.

Cell-to-cell cross-talk, such as that described here, may represent a series of mechanisms by which viruses alter a particular microenvironment, and change the cellular milieu to one that is conducive to viral replication and maintenance. We need to understand these mechanisms, so as to manipulate that milieu to favour antiviral therapies. The aim will be not only to decrease the production of HIV-1 but also to destroy the cellular reservoirs that have so far kept us from being able to eradicate HIV-1 infection¹⁴.

Roger J. Pomerantz is in the Division of Infectious Diseases and Environmental Medicine, Center for Human Virology and Biodefense, Thomas Jefferson University, 1020 Locust Street, Suite 329, Philadelphia, Pennsylvania 19107, USA.
e-mail: roger.j.pomerantz@jefferson.edu

- Swingler, S. *et al.* *Nature* **424**, 213–219 (2003).
- Unutmaz, D., KewalRamani, V. N., Marmon, S. & Littman, D. R. *J. Exp. Med.* **189**, 1735–1746 (1999).
- Ducrey-Rundquist, O., Guyader, M. & Trono, D. *J. Virol.* **76**, 9103–9111 (2002).
- Zhang, Z. *et al.* *Science* **286**, 1353–1357 (1999).
- Wong, J. K. *Science* **278**, 1291–1295 (1997).
- Finzi, D. *Science* **278**, 1295–1300 (1997).
- Kirchhoff, F., Greenough, T. C., Brettler, D. B., Sullivan, J. L. & Desrosiers, R. C. *N. Engl. J. Med.* **332**, 228–232 (1995).
- Kestler, H. W. *Cell* **65**, 651–662 (1991).
- van Kooten, C. & Banchemare, J. *J. Leukoc. Biol.* **67**, 2–17 (2000).
- Pierson, T. C. *et al.* *J. Virol.* **76**, 8518–8531 (2002).
- Wu, Y. & Marsh, J. W. *Science* **293**, 1503–1506 (2001).
- Pomerantz, R. J. *AIDS* **15**, 1201–1211 (2001).
- Swingler, S. *et al.* *Nature Med.* **5**, 997–1003 (1999).
- Kulkosky, J. *et al.* *J. Infect. Dis.* **186**, 1403–1411 (2002).

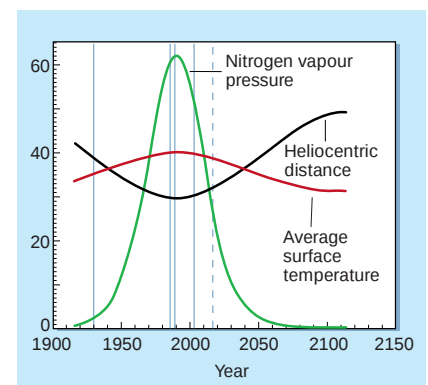


Figure 1 Pluto variations. As Pluto's distance from the Sun changes (black curve, in astronomical units), its average surface temperature should vary roughly inversely as the square root of the distance (red curve, in kelvin). The vapour pressure of surface nitrogen ice changes even more markedly (green curve, in microbar), and should have been decreasing between 1988 and 2002. In fact, Elliot *et al.*¹ and Sicardy *et al.*² show that Pluto's atmosphere increased in size over this interval, but this could possibly be attributed to a time-lag in surface-layer cooling as the planet moves away from the Sun. The vertical blue lines indicate, left to right, the 1930 discovery of Pluto; the occultation observations in 1985, 1988 and 2002, the last of these by Elliot *et al.*¹ and Sicardy *et al.*²; and (dashed) the earliest possible arrival of a spacecraft at Pluto.

through a slight distortion of the spacecraft's radio link to Earth, that Triton too has a tenuous atmosphere, with a surface pressure of perhaps 10 to 20 microbar (atmospheric pressure on Earth is one bar).

At the time of the Triton encounter, Pluto was actually slightly closer to the Sun than Neptune was. But Pluto always maintains a healthy distance from Neptune, because their orbits are synchronized. So a spacecraft visit to Pluto would require a specially planned trajectory and a lengthy trip, rather than any straightforward extension of an existing mission. And as Pluto's diameter is only some 2,400 km, substantially smaller than the Moon's, is it worth the trip?

In 1988, just one year before the Voyager 2 encounter with Triton, distortion of light from an occulted star revealed that Pluto had a tenuous atmosphere, to which Triton's is somewhat similar. In fact, the earliest evidence for Pluto's atmosphere seems to have been gathered in 1985 by the Israeli astronomer Noah Brosch: observing in the Negev desert, he saw a gradual dimming of starlight caused by refraction in Pluto's atmosphere, rather than the knife-edge drop expected when an atmosphere-less planet (such as the Moon) occults a star. Brosch's data were so unexpected that he was widely disbelieved until the case was clinched by observations of the 1988 occultation by

Planetary science

Pluto's atmospheric surprise

William Hubbard

Last year, for the first time in 14 years, an alignment of stars with Pluto created an opportunity to observe the atmosphere of this most remote of planets. Though tenuous, the atmosphere has, remarkably, expanded.

Pluto is the only planet in the Solar System still unvisited by spacecraft.

But an appropriate conjunction of stars with the planet can make some Earth-bound observations possible. In particular, Pluto's thin atmosphere (exerting a surface pressure roughly a million times lower than that at the Earth's surface) comes into view as the planet passes in front of, or 'occults', a bright star. On pages 165 and 168 of this issue, Elliot *et al.*¹ and Sicardy *et al.*² report the first new data on Pluto's atmosphere from stellar occultation in more than a decade. The observations are timely: Pluto's orbit over the next few years offers an opportunity to learn more about this planet, at a time when technological developments make it feasible to consider a mission to it. But both time and money are in short supply.

As Pluto is the most distant of the nine planets, its 248-year orbit around the Sun is lengthy compared with human timescales. It is also the most eccentric: Pluto's heliocentric distance varies between 30 and 50 astronomical units (one astronomical unit, or AU, is equivalent to the mean distance between the Earth and the Sun). In accordance with Kepler's second law of planetary motion, Pluto spends most time in each orbit at almost its maximum distance from the Sun and the Earth. But, in 1989, Pluto coincidentally made its closest approach to the Sun (called its perihelion) during a period of heavy spacecraft exploration and in precisely the year that the spacecraft Voyager 2 had its final encounter with Neptune and Neptune's icy satellite Triton. Triton seems to be a very similar body to Pluto and is perhaps a close relative³. The Voyager 2 encounter revealed,

several telescopes, including NASA's Kuiper Airborne Observatory⁴.

What process could produce microbar atmospheres on tiny objects such as Triton and Pluto? Triton and Pluto are thought to belong to the Kuiper belt — a ring of icy bodies beyond the orbit of Neptune — and so bear a generic relation to short-period comets whose orbits pass through the belt. Like comets, they seem to be able to generate fluctuating atmospheres in response to variable solar heating of their icy surface layers. The difference, though, is that Triton and Pluto are just massive enough to retain the gas as a bound atmosphere, whereas comets produce an escaping envelope of gas, known as a coma. And because the solid surface layers of Triton and Pluto never get warmer than about 40 kelvin, any gas that erupts must be much more volatile than the water vapour thought to power the spectacular tails of those comets whose orbits bring them within a few AU of the Sun.

By a process of elimination, a likely candidate for the main gas in Pluto's atmosphere is the volatile molecule nitrogen. As Fig. 1 shows, its vapour pressure could build to a maximum at perihelion well in excess of the pressures measured in Pluto's atmosphere. But because the vapour pressure of nitrogen decreases strongly as the surface temperature falls, the atmosphere ought to collapse as Pluto's surface cools during its retreat from the Sun over the next few decades.

Following the initial detections of Pluto's atmosphere in 1985 and 1988, there was an unproductive 14-year period during which astronomers sought, to no avail, either to place portable telescopes in an occasional narrow path of Pluto's starlight shadow on the Earth, or to predict feasible occultation opportunities at fixed observatories. There were, typically, one or more aborted campaigns every few years or so. The forecast shadow paths invariably turned out to be off Earth or otherwise inaccessible. The shutdown of the Kuiper Airborne Observatory did not help. Planetary scientists have now become strong advocates for a spacecraft mission to Pluto while the planet is still warm enough to exhibit the volatile molecules embedded in its surface layers.

As discussed by Elliot *et al.*¹ and Sicardy *et al.*², Pluto's atmosphere has, unexpectedly, expanded rather than contracted over the past 14 years, but temperature variations in planetary surface layers do typically lag somewhat behind solar heating variations. In the long run, cooling and atmospheric contraction are inevitable. To understand what is going on, we need a spacecraft mission to Pluto: an occultation measurement of the atmosphere at radio wavelengths, together with ultraviolet measurements of a solar occultation, could tell us much more — and both of these require an antenna or telescope to aim past Pluto back towards the

Earth or the Sun. Meanwhile, possible light extinction caused by particles in Pluto's atmosphere (reported by Elliot *et al.*) may be bad news for detailed stellar-occultation studies of the atmosphere, as standard analysis methods assume that there is no loss of photons.

The first of NASA's New Frontiers line of space missions will be a Pluto–Kuiper-belt mission called New Horizons⁵. The mission has a planned launch in 2006 and a Pluto flyby about ten years later (Fig. 1). But New Horizons is in trouble because of unexpected extra costs. Similarly, NASA's Planetary Astronomy programme, which has supported much of the Pluto occultation work in the United States, had to cancel

some of the ground stations that could have provided more data. Like mountaineers preparing an assault on Mount Everest, scientists are closely watching the calendar and their budgets, hoping to reach their objective before the snows come.

William Hubbard is at the Lunar and Planetary Laboratory, University of Arizona, Tucson, Arizona 85721-0092, USA.
e-mail: hubbard@lpl.arizona.edu

1. Elliot, J. L. *et al.* *Nature* **424**, 165–168 (2003).
2. Sicardy, B. *et al.* *Nature* **424**, 168–170 (2003).
3. McKinnon, W. B., Lunine, J. I. & Banfield, D. in *Neptune and Triton* (ed. Cruikshank, D. P.) 807–877 (Univ. Arizona Press, Tucson, 1995).
4. Binzel, R. P. & Hubbard, W. B. in *Pluto and Charon* (eds Stern, S. A. & Tholen, D. J.) 85–102 (Univ. Arizona Press, Tucson, 1997).
5. <http://pluto.jhuapl.edu>

Cell biology

The molecules that make muscle

Donald Gullberg

The role of integrin proteins in the formation of skeletal muscle has been hotly debated, as studies of whole animals and of cultured cells have yielded conflicting results. The controversy may now be resolved.

Every conscious action we perform requires skeletal muscle. Without it, we wouldn't be able to run for the bus, lift a pipette or turn the pages of this journal. So called because it attaches to the skeleton, this type of muscle has been studied for centuries; incredibly, the idea that it works by contracting can be traced back more than two millennia. As a consequence, a great deal is known about the structure of skeletal muscle and about how it functions. The step-by-step process by which it forms has also been well studied, at least at the cellular level. Nowadays, however, the goal of much research effort is to identify the molecules that are involved in specific biological processes — and the molecular mechanisms underlying muscle formation have long eluded researchers. That now looks set to change: writing in *Developmental Cell*, Schwander and colleagues¹ report evidence supporting a key role for members of the integrin family of cell-adhesion proteins.

Integrins are proteins, found in multicellular organisms, that are embedded in the plasma membrane. They enable cells to adhere to the matrix of molecules in which they find themselves, and they also bind to various intracellular proteins, thereby connecting the cell exterior and interior. Each integrin is made up of two proteins, one from the α subfamily and one from the β subfamily, and they are involved in a variety of biological functions, such as cell proliferation, differentiation and migration, and the prevention of cell death.

Many of these functions have been inferred from animal studies, in which

individual integrin proteins were mutated and the effects studied. But studies of integrins *in vitro* have yielded conflicting data — as have studies involving different species. One example concerns skeletal muscle formation (myogenesis; Fig. 1)². During this process, individual muscle precursor cells first become 'destined' to form muscle; at this stage they are called myoblasts. The myoblasts then migrate to the site of future muscle formation, where they fuse to produce skeletal muscle cells — myotubes — that have multiple nuclei. The smallest contractile unit in a mature myotube is the sarcomere, a regular assembly of filaments made up of the myosin and actin proteins. The sarcomere is assembled as the myotubes mature into muscle fibres.

So what is the confusion over integrin's role in this process? Early studies in the 1980s showed that blocking integrins containing a $\beta 1$ subunit on chick myoblasts *in vitro* inhibited fusion³. Later work with fruitfly myoblasts lacking the invertebrate variant of the vertebrate integrin $\beta 1$ chain pointed to a role for the protein in sarcomere assembly as well⁴. But attempts to show convincingly that $\beta 1$ integrins are involved in these events in mice have failed. One of the problems is that mouse embryos engineered to lack functional integrins throughout the body die long before the effects of the deficiency on muscle function can be studied^{5–7}.

To overcome these problems, cells lacking both functional gene copies for specific integrins have been injected into wild-type early embryos, thereby generating chimaeric mice (in which only some of the cells lack

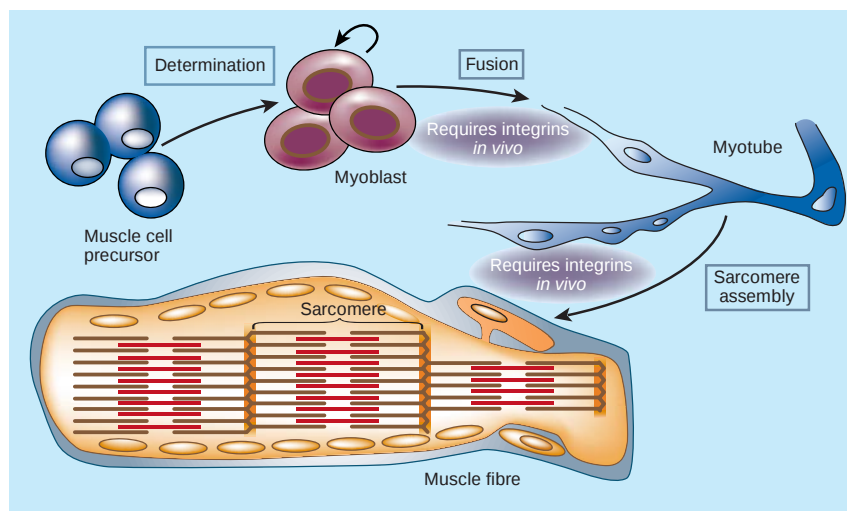


Figure 1 Muscle formation. The process begins with muscle cell precursors becoming 'destined', or 'determined', to form muscle, at which point they are called myoblasts. The myoblasts then proliferate and migrate to the site of future muscle formation, where they fuse to produce myotubes. As myotubes mature, the sarcomere develops; this array of myosin and actin fibres forms the contractile unit of muscle. Schwander *et al.*¹ have found that $\beta 1$ integrin is necessary for myoblast fusion and for sarcomere assembly in mice.

integrins)⁷⁻⁹. The results, however, have not been easy to interpret. One might think that the effects of integrin mutations on the formation of muscle precursor cells and on myoblast migration would be readily amenable to analysis. The problem here is that there are still wild-type cells present in chimaeric mice, and these, through cell-cell interactions, could rescue defective precursor formation or cell migration (for instance, it has been suggested that integrin-deficient cells could piggy-back on wild-type cells during migration)⁷. Once myoblasts have fused, it is even more difficult to evaluate the effects of integrin mutations, because wild-type cells will be able to 'supply' integrin to myotubes.

One way around these difficulties *in vivo* has been to isolate myoblasts or satellite cells (skeletal-muscle stem cells) from chimaeric muscle, culture just the integrin-deficient cells, and use them to study myoblast fusion, differentiation and sarcomere formation *in vitro*. But these studies⁸⁻¹² have failed to support earlier findings of a requirement for integrins (and for another family of cell-adhesion receptors, the cadherins) in myoblast fusion and sarcomere assembly.

Schwander *et al.*¹ now propose that this strategy, too, could be flawed: they suggest that, when cells are cultured for long periods, compensatory mechanisms that obscure the original characteristics are set into play. This interesting hypothesis remains to be tested, but in the meantime Schwander *et al.* have used a different technique (the Cre/Lox system) to inactivate the $\beta 1$ integrin gene specifically in muscle cells in mouse embryos.

The authors found that these mutant animals died at birth, with highly under-developed muscles. Analysis of the muscles

showed that $\beta 1$ integrins are not needed for myoblast proliferation or migration. But the authors' studies of cultured $\beta 1$ -deficient cells, taken from the mutant embryos, suggested that these proteins are necessary for a late event in myoblast fusion. *In vitro* data also showed that $\beta 1$ integrins are needed for correct sarcomere assembly in myotubes. There are apparently $\beta 1$ -integrin-independent mechanisms that allow fusion and sarcomere assembly to a limited extent *in vitro*, but *in vivo* these mechanisms are not sufficient to ensure correct muscle formation.

Previous data have unequivocally shown that integrins are needed for muscles to attach to the body wall in fruitflies¹³. The same appears to be true of vertebrates, as a lack of $\alpha 7 \beta 1$ integrin in mice causes a form of muscular dystrophy that primarily affects the junctions between muscle and tendons¹⁴. It now seems that another conserved function of integrins is in myoblast fusion and sarcomere assembly. The authors suggest that integrin $\beta 1$ might be cooperating with a protein called CD9 in the events leading up to fusion, as CD9 fails to localize properly in the plasma membrane in the absence of $\beta 1$ integrins. This fits with the previous finding that CD9 is implicated in both myoblast-myoblast fusion and sperm-egg fusion. It will be important to determine exactly how it might be working with $\beta 1$ integrins in myoblast fusion. The role of $\beta 1$ integrins in sarcomere assembly most probably involves attaching the sarcomeres to the plasma membrane.

A goal for the future will be to sort out which members of the α -integrin subfamily pair up with $\beta 1$ integrin to support fusion and sarcomeric assembly; good candidates



100 YEARS AGO

The following argument may be of interest to your readers, if, as I suspect, it has never been thus formulated before. In order to account for the origin of species, we must assume that the tendency of an individual to vary is handed down to future generations by appropriate modifications in the transmitted germ-plasma. But, unless we believe in the inheritance of acquired characters (for which we have no certain evidence), the tendency of the first individual to vary can only have become manifest if it had originated in a modification of its own parents' germ-plasma: otherwise that tendency could not have been inherited. Leaving out of account the play of changing external conditions, we are thus forced to regard the variability of individuals as the result of variations in the parental germ-plasma. The problem is, how are such variations produced?
From *Nature* 9 July 1903.

50 YEARS AGO

The scientist's responsibilities. It is widely maintained that "the scientist establishes facts: other sections of society make what use of them they will, and that use is to a great extent unpredictable". I think this is a serious under-estimate of the scientist's vocation. Admittedly he works within the framework of objectively ascertained facts, and admittedly there is a big element of unpredictability in the implications of any new discovery. But Nature is so complicated and manifold that those aspects or facts that a generation's work can hope to reveal are only a minute proportion of what is waiting there for future generations. Some of the spectacular advances of recent years have induced a sort of intellectual arrogance among scientists, an over-estimation of the extent of our understanding of Nature which leads them to under-estimate the enormous choice of pathways of experimentation and observation which lies open to us. This leads to a very routine approach to experimentation, and one which is much subject to crazes. Some fields of study are worked to exhaustion and others are neglected. The desire of those who finance research for immediate practical results is only partly to blame for this. Scientists must come to feel much more responsibility for what facts are ascertained, that is, for the direction taken by the progress of science.
From *Nature* 11 July 1953.

are the $\alpha 4$ and $\alpha 5$ integrin chains. It will also be interesting to see whether different integrin dimers are needed at different steps in the process. A lesson from Schwander and colleagues' work is that these experiments must be carefully designed to take into account the compensatory effects of other integrins and of non-integrin-dependent compensatory systems at work in muscle cells.

Donald Gullberg is in the Department of Medical Biochemistry and Microbiology, Biomedical Centre, Husargatan 3, SE-751 23 Uppsala, Sweden.
e-mail: donald.gullberg@imbim.uu.se

1. Schwander, M. et al. *Dev. Cell* **4**, 673–685 (2003).

2. Gullberg, D., Tiger, C.-F., Lohikangas, L. & Velling, T. *Front. Biosci.* **3**, 1028–1039 (1998).
3. Menko, A. S. & Boettiger, D. *Cell* **51**, 51–57 (1987).
4. Volk, T., Fessler, L. I. & Fessler, J. H. *Cell* **63**, 525–536 (1991).
5. Yang, J. T., Rayburn, H. & Hynes, R. O. *Development* **119**, 1093–1105 (1993).
6. Yang, J. T., Rayburn, H. & Hynes, R. O. *Development* **121**, 549–560 (1995).
7. Fässler, R. & Meyer, M. *Genes Dev.* **9**, 1896–1908 (1995).
8. Yang, J. T. et al. *J. Cell Biol.* **135**, 829–835 (1996).
9. Taverna, D. et al. *J. Cell Biol.* **143**, 849–859 (1998).
10. Charlton, C. A., Mohler, W. A., Radice, G. L., Hynes, R. O. & Blau, H. M. *J. Cell Biol.* **138**, 331–336 (1997).
11. Hollnagel, A., Grund, C., Franke, W. W. & Arnold, H. H. *Mol. Cell. Biol.* **22**, 4760–4770 (2002).
12. Hirsch, E., Lohikangas, L., Gullberg, D., Johansson, S. & Fässler, R. *J. Cell Sci.* **111**, 2397–2409 (1998).
13. Brown, N. H. *BioEssays* **15**, 383–390 (1993).
14. Mayer, U. et al. *Nature Genet.* **17**, 318–323 (1997).

Tumour suppressors

Timing will tell

Anton Berns

A sophisticated model of the effects of mutating the retinoblastoma gene reveals a window of opportunity in which cells can be released from the normal control of proliferation, before back-up mechanisms take effect.

How much flexibility is there in the molecular circuits that regulate the functions of cells? If some genes are inactivated, do back-up systems restore the status quo? On page 223 of this issue, Jacks and co-workers¹ address this question with regard to the retinoblastoma gene, *Rb*, which works in a key cellular signalling pathway that protects organisms against cancer².

The *Rb* gene is known to put the brakes on cell division, controlling cell numbers and thereby helping to prevent tumours from forming. Mutations that inactivate this 'tumour suppressor' gene are often seen in human cancers. One way in which researchers try to learn more about how the inactivation of particular genes contributes to cancer development is to mutate them in mice, and see what happens. But a problem with such mouse models is that they generally involve mutating the gene of interest in the germline, the tissue that produces eggs and sperm. This means that the gene is inactivated in all cells, and throughout the lifetime, of offspring — at least if the loss of gene function does not lead to the death of embryos (as happens with *Rb*). By contrast, sporadic human cancers involve the progressive accumulation of tumour-predisposing mutations in just a few cells.

'Conditional knockout' mice allow much more sophisticated cancer modelling³: these animals are engineered such that genes can be inactivated in a spatially and temporally controlled manner. And they have another benefit. They enable biologists to find out whether — and, more particularly, when — the loss of a particular gene can be compensated for by the activity of other genes. Compensation

could occur directly, if the back-up genes are simply upregulated or downregulated as appropriate as a direct result of *Rb* loss. Or it could occur indirectly, by mechanisms such as chromatin remodelling, which alters the configuration of DNA and hence can affect the expression of many genes. Such indirect mechanisms come into play abundantly during development — a distinct chromatin configuration is often fixed during development in a cell-type-specific fashion, and thereby imposed on all daughter cells. By contrast, the direct mechanisms work more or less independently of cellular history. So, conditional knockouts allow one to ask whether direct compensatory mechanisms are involved, and, if so, how quickly they respond. In other words, timing tells.

So, does the timing of *Rb* inactivation make a difference? Jacks and colleagues¹ addressed this question in their studies of cultured mouse embryonic fibroblast cells (MEFs), derived either from animals in which *Rb* was germline-mutated, or from mice in which *Rb* could be conditionally ('acutely') inactivated. They noted some striking differences between the two types of cell.

The authors saw no variation in the proliferation properties of exponentially dividing, 'early passage' MEFs (cells that had just been taken from mice and cultured). But the germline-mutated cells did show a prominent increase in their levels of p107 (a close relative of the *Rb* protein) and of other regulators of the cell-division cycle, such as cyclin E and p19^{Arf}. By contrast, in those MEFs in which *Rb* was acutely inactivated during culture, the levels of these proteins were low at first. But, after several passages (that is, after a few cells had been taken from the initial culture dish

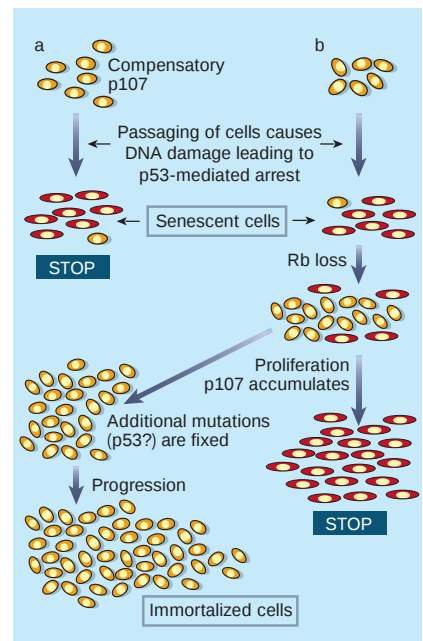


Figure 1 Timing is everything. Jacks and colleagues¹ have studied mouse embryonic fibroblast cells (MEFs) taken from a, animals that lacked the retinoblastoma (*Rb*) gene from the start, or b, mice in which *Rb* could be mutated at a specific time. a, Here, levels of the p107 protein (and others) have increased to compensate for the loss of *Rb*. Repeated passaging of these cells causes DNA damage, which is sensed by the p53 protein, causing the cells to stop dividing. They thereby enter a senescent state (represented by large, flat cells), which p107 helps to maintain. b, These cells express *Rb* at first, and so can enter the senescent state as normal after repeated passaging. But after *Rb* has been experimentally deleted, many of the cells start dividing again (represented by smaller cells). This provides a time window during which they could accumulate more tumour-predisposing mutations (in, for instance, the p53 pathway). Alternatively, the increase in p107 levels that occurs could cause the cells to become senescent once again.

and plated on a new one, allowed to multiply again, and so on), the levels were similar to those in germline-mutated MEFs (T. Jacks, personal communication). This implies that the expression of at least these genes could be recalibrated fairly soon after *Rb* loss.

Very different effects were seen, however, when *Rb* was inactivated in MEFs made quiescent (non-dividing because of a lack of growth factors in the medium), or senescent (non-dividing as a result of repeated passaging or overexpression of the Ras^{V12} protein). Jacks and colleagues found that, in quiescent cells, acute deletion of *Rb* resulted in a burst of DNA synthesis, indicating that the cells had re-entered the cell cycle. This did not occur in cells carrying a germline *Rb* mutation, hinting that it was the specific timing of

Rb inactivation that triggered cell-cycle re-entry. The authors went on to show that p107 and p19^{Arf} both contributed to maintenance of the quiescent state in the germline-inactivated cells, suggesting that these proteins help to compensate for the loss of *Rb* in these circumstances.

The effects on senescent cells were even more dramatic (Fig. 1). Senescence is a condition of proliferation arrest from which cells rarely escape. Senescent cells are usually large, flat and metabolically active, yet do not divide; they express a set of characteristic 'marker' proteins; and they upregulate cell-cycle inhibitors such as p19^{Arf} and p16^{Ink4a} (ref. 4). Senescence can prevent mammalian cells from becoming malignant⁵, and may influence the efficacy of anticancer treatments⁶. Jacks and colleagues found that acute mutation of *Rb* in senescent MEFs forced a significant proportion of the cells back into the division cycle (as judged by the incorporation of a division-specific dye, BrdU, and by the rounding-up of large, flat cells). Moreover, many of the cells continued to divide for a number of passages.

Senescence in MEFs is probably caused by the accumulation of damaged DNA⁷. Usually, DNA damage is sensed by a signalling pathway involving another tumour suppressor, p53, which then halts cell division. Previous work⁸ showed that when p53 was inhibited in senescent cells, this too caused them to round up (although most then died). An earlier study⁹ also found that either inhibition of the p53 pathway, or a lack of at least two *Rb*-related proteins (also known as pocket proteins), can confer immortality — the ability to continue dividing — on MEFs. And the fact that *Rb*-deficient, senescent MEFs re-enter the cell cycle¹ is consistent with the observation that p53-mediated cell-cycle arrest requires a certain level of pocket proteins bound to the E2F protein¹⁰. All of this implies that acute, temporary loss of *Rb* might relieve a p53-imposed block in senescent MEFs, enabling them to continue proliferating with damaged DNA. That can't happen when *Rb* is inactivated in the germline, because p107 compensates.

There is increasing evidence that regulators of the p53 and *Rb* signalling pathways, such as p19^{Arf} and p16^{Ink4a}, respectively, have different effects on the induction of senescence in human and mouse fibroblasts. This could well relate to the propensity of MEFs to accumulate DNA damage under cell-culture conditions — putting the emphasis on the p53 pathway in these cells.

Although this might make MEFs rather atypical, Jacks and colleagues' observations¹ do illuminate a mechanism with broad ramifications. The instant imbalance brought about by acute loss of tumour-suppressor function can 'kick' cells into the division cycle, or overcome other mechanisms that

keep cells in check, thereby providing a window of opportunity for incipient cancer cells to progress to a more malignant state. The observation that it requires several passages before MEF cultures revived from senescence by *Rb* inactivation can re-enter senescence¹ supports the idea that it takes time to activate back-up systems. That this window of opportunity can present a real threat might be illustrated by the fact that 12 of the 14 senescent cultures that began dividing upon *Rb* inactivation subsequently became immortal — quite possibly as a result of acquired mutations in the p53 pathway.

One other point is worth noting. The data reported by Jacks and colleagues do not provide evidence for enduring differences in gene-expression profiles between cells that lack *Rb* from the start, and those in which *Rb* is acutely deleted. But that does not necessarily mean that there are no such

differences, fixed early in development by mechanisms such as chromatin remodeling¹¹. Time will undoubtedly tell.

Anton Berns is in the Division of Molecular Genetics, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands.
e-mail: a.berns@nki.nl

1. Sage, J., Miller, A. L., Pérez-Mancera, P. A., Wysocki, J. M. & Jacks, T. *Nature* **424**, 223–228 (2003).
2. Chau, B. N. & Wang, J. Y. *Nature Rev. Cancer* **3**, 130–138 (2003).
3. Vooijs, M. & Berns, A. *Oncogene* **18**, 5293–5303 (1999).
4. Sherr, C. J. *Nature Rev. Mol. Cell Biol.* **2**, 731–737 (2001).
5. Campisi, J. *Trends Cell Biol.* **11**, S27–S31 (2001).
6. Schmitt, C. A. et al. *Cell* **109**, 335–346 (2002).
7. Parrinello, S. et al. *Nature Cell Biol.* doi:10.1038/ncb1024 (2003).
8. Dirac, A. M. & Bernards, R. *J. Biol. Chem.* **278**, 11731–11734 (2003).
9. Dannenberg, J. H., van Rossum, A., Schuijff, L. & te Riele, H. *Genes Dev.* **14**, 3051–3064 (2000).
10. Rowland, B. D. et al. *Cancer Cell* **2**, 55–65 (2002).
11. Jaenisch, R. & Bird, A. *Nature Genet.* **33** (Suppl.), 245–254 (2003).

Oceanography

Stirring times in the Atlantic

Bob Dickson

Set against certain other oceanographic phenomena in the northwest Atlantic, a class of brief and localized events that occur in the Irminger Sea may seem inconsequential. Not so, however.

On page 152 of this issue, Pickart et al.¹ describe the winter incidence and climatology of so-called 'tip-jet events' off southern Greenland, their association with storm centres tracking north-eastward into the Denmark Strait, and their effect in producing episodes of deep convective exchange in the Irminger Sea to the southeast of Greenland. That tip jets should have this effect is no surprise in itself. Although they are brief, localized, intermittent and strongly seasonal, such events are evidently associated with an intense heat flux from the ocean and a strong wind stress curl (the 'twisting' tendency in the wind stress field; Fig. 1, overleaf) that are important drivers of ocean convection. The surprise comes in the suggestion that sufficient 'Labrador Sea Water' might be produced in the winter Irminger Sea to rival that formed in the Labrador Sea itself¹.

The Labrador Sea, which lies to the south and west of Greenland, is of particular climatic importance. It is at the headwaters of the Atlantic Ocean's thermohaline 'conveyor', and is the basin where a wide range of water masses, including the output from Arctic and subarctic seas, are received, stored, transformed and discharged to form the deep western boundary current — the abyssal limb of the Atlantic conveyor. The loss of heat and buoyancy under the chill northwesterly airflow from Arctic Canada in winter has also made the Labrador Sea one of

the main global centres of deep convective exchange. It is thus a site where one of the principal vertically homogeneous 'mode waters' of the ocean — Labrador Sea Water (LSW) — is formed, and where the signal of climate change can be transferred to great ocean depths.

Over the past three or four decades, a broad range of upstream influences have come together to drive ocean variability in the Labrador Sea on multi-decadal timescales and at all ocean depths. The most dramatic instance has been the observed transfer of 'freshening' from subarctic seas to the entire water column of the northwest Atlantic.

In the upper ocean, this occurred through the relatively fresh, relatively shallow, mix of Pacific water, Arctic ice-melt and Arctic river discharge that passes south to the Atlantic on either side of Greenland. There is clear, if indirect (proxy), evidence of a 40-year increasing trend in the amount of fresh near-surface waters passing south from the subarctic around the western margins of the Labrador Sea². In the Labrador Sea itself, these cool, fresh surface conditions were picked up and mixed to intermediate depths of more than 2,000 m by convective exchanges of intensifying — and ultimately of record — vigour, stimulated by the chill northwesterly airflow sweeping out of Arctic Canada at that time. By 1994, the layer of convectively formed LSW was fresher, colder, deeper and denser than at any other time in

the history of measurements there. Finally, in the lower layers, a long-term and broad-scale freshening of the seas north of Iceland^{3–5} was communicated south to the deep ocean by the two dense overflows that descend from their spillways across the Greenland–Scotland Ridge to ventilate the deep and abyssal Atlantic; this entire system of overflow and entrainment has freshened steadily over the past four decades⁶.

Many of these individual changes are beyond the range of our past experience. Most have their origins in the long-period amplification of the leading mode of wintertime atmospheric forcing in these waters, the North Atlantic Oscillation (NAO). By the early 1990s, the NAO had evolved to its most extreme positive state in the 175 years of instrumental recording and possibly much longer⁷ (the positive state of the NAO is one in which an intensified winter pressure gradient between Iceland and the Azores drives strong mid-latitude westerlies). As a result, between 1966 and 1994, in what we believe to be the largest change in the modern instrumental oceanographic record, a sustained climatic change imposed on Arctic and subarctic seas was transferred to all layers — shallow, intermediate, deep and abyssal — of the northwest Atlantic.

Set against events of such monumental scale, the new mode of 'LSW' production reported by Pickart *et al.*¹ may seem small beer. But not so. The mechanism they describe is certainly strongly intermittent, occurring only in winter, and with a ten-day period and three-day average duration. Nonetheless, when Pickart *et al.* add up the numbers, they conclude that the heat flux and strong wind stress curl associated with these tip-jet events drive localized convection of sufficient depth (up to 2,000 m) and persistence to produce significant quantities of 'LSW' outside the Labrador basin.

Pickart *et al.* appeal more to ocean modelling than to ocean observations to support their suggestion. And in fact, as long as our knowledge of change in these remote waters is so heavily based on occasional (usually non-winter) surveys, and while the signal of decadal change in classic LSW is still passing through these waters, it may prove difficult to assess the importance of convection in the Irminger Sea more directly. However, there are good reasons for persisting in the attempt.

The first is the central role of these waters in the Earth's climate system. Through the northwest Atlantic 'junction box' pass all of the deep and bottom waters that collectively form and drive the abyssal limb of the Atlantic meridional overturning circulation. And around it pass the freshwater flows from the Arctic that, in model experiments, are implicated in slowing that circulation down. Furthermore, a sizeable body of theory^{8–11} now suggests that the vertical circulation too may help determine the rate of the

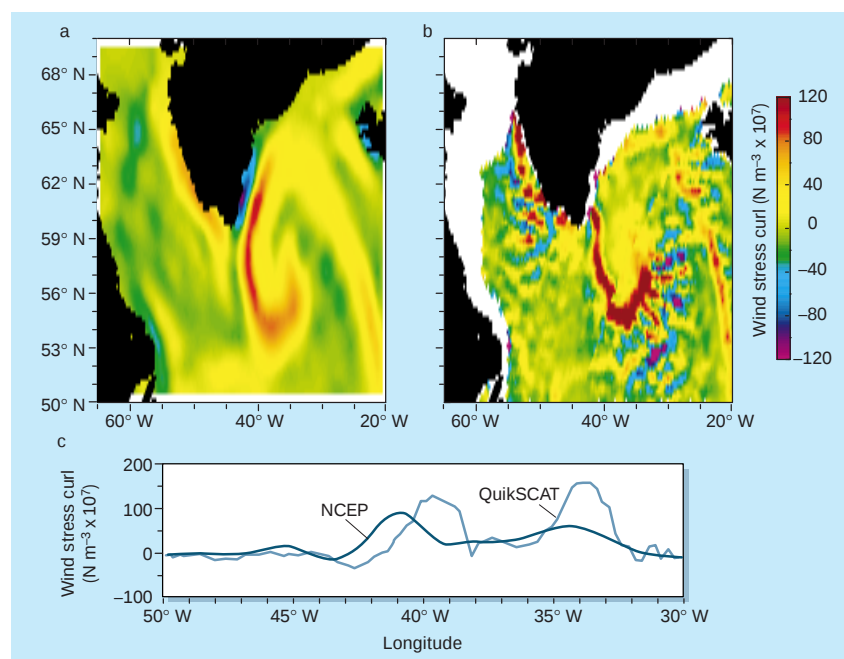


Figure 1 The finer scales of satellite sensing. A conclusion to be drawn from the study by Pickart *et al.*¹ is that fine-scale ocean forcings may influence climate. As shown in these images, satellite-borne instruments now provide a remarkable ability to provide synoptic coverage of winds at the 25-km scale over water. This new capability is demonstrated for 12 January 2001, as a wind jet develops at the southern tip of Greenland (the central black image), in association with a rapidly moving cyclonic centre over the Irminger Sea to the southeast. **a**, Conventional analysis of wind stress curl distributions from the National Centers for Environmental Prediction (NCEP). **b**, The higher resolution available from the radar scatterometer on the QuikSCAT satellite¹⁵. This image reveals intense, small-scale features and fronts, crucial to the field of wind stress curl that — as the twisting tendency of the wind stress field — drives ocean upwelling and convection. **c**, Graphical depiction of the large differences in the recorded amplitude of wind stress curl between the two techniques, for latitude 56.5° N. (Analysis kindly provided by Dudley Chelton, Oregon State Univ.)

overturning circulation by setting the density of the water column. Predictions of the ocean's role in climate therefore depend on portraying the convective processes of this area with appropriate realism.

It is possible, too, that the relative importance of the phenomenon described by Pickart *et al.* may be growing. In recent winters, a slow eastward displacement of the NAO pattern has reduced the incidence of chill northwesterly winds over the Labrador Sea while increasing the northwesterly airflow from south Greenland across the Irminger Sea. Although the effects of greenhouse gases have produced this kind of shift in models¹², Hurrell *et al.*¹³ conclude that it is more likely to be the intrinsic result of a more NAO-positive climate. Its significance is not only that the warmth of recent winters over the Labrador Sea has, since 1996, brought a dwindling to extinction of deep, dense LSW production¹⁴, but also that the northwesterly winds over southeast Greenland seem a necessary stimulus for tip-jet forcing of convective events in the Irminger Sea, as described by Pickart and colleagues.

The climatic influence of those events will be felt through the conversion of such intermittent and small-scale events in the

atmosphere into a sustained ocean response. The challenge now is to observe that response directly, and the necessary new ocean-profiling sensor systems have already been deployed for that purpose off southeast Greenland. ■

Bob Dickson is in the Environmental Quality Group, Centre for Environment, Fisheries and Aquaculture Science (CEFAS), Lowestoft, Suffolk NR33 0HT, UK.

e-mail: r.r.dickson@cefasc.co.uk

- Pickart, R. S., Spall, M. A., Ribergaard, M. H., Moore, G. W. K. & Milliff, R. F. *Nature* **424**, 152–156 (2003).
- Dickson, R. R., Curry, R. & Yashayaev, I. *Phil. Trans. R. Soc. Lond. A* (in the press).
- Verdun, J. & Quadfasel, D. in *European Sub-Polar Ocean Prog. II, Final Sci. Rep.* (ed. Jansen, E.) (Univ. Bergen, 1999).
- Blindheim, J. *et al. Deep-Sea Res.* **147**, 655–680 (2000).
- Hansen, B., Turrell, W. R. & Østerhus, S. *Nature* **411**, 927–930 (2001).
- Dickson, R. R. *et al. Nature* **416**, 832–837 (2002).
- Jones, P. D. *ICES Mar. Sci. Symp. Ser.* (in the press).
- Rooth, C. *Progr. Oceanogr.* **11**, 131–149 (1982).
- Rahmstorf, S. *Clim. Dyn.* **12**, 799–811 (1996).
- Marotzke, J. & Klinger, B. A. *J. Phys. Oceanogr.* **30**, 955–969 (2000).
- Thorpe, R. B., Gregory, J. M., Johns, T. C., Wood, R. A. & Mitchell, J. F. B. *J. Clim.* **14**, 3102–3115 (2001).
- Ulrich, U. & Christoph, M. *Clim. Dyn.* **15**, 551–559 (1999).
- Hurrell, J. W., Kushnir, Y., Ottersen, G. & Visbeck, M. in *The North Atlantic Oscillation: Climatic Significance and Environmental Impact* (eds Hurrell, J. W. *et al.*) 1–35 (Am. Geophys. Un., Washington DC, 2003).
- Yashayaev, I., Lazier, J. R. N. & Clarke, R. A. *ICES Mar. Sci. Symp. Ser.* (in the press).
- Draper, D. W. & Long, D. G. *J. Geophys. Res.* **107**, 3212 (2002).

Nanotechnology

Nanotubes get sorted

Science doi:10.1126/science.1086534 (2003)¹
J. Phys. Chem. doi:10.1021/jp027664a (2003)²

Researchers wanting to use carbon nanotubes as components for nanoscale electronics face the problem that standard synthetic methods produce a combination of metallic and semiconducting tubes. The differing electronic properties are a consequence of only very subtle differences in molecular structure, making the two types hard to separate. But now there is a choice of two techniques for doing this.

Ralph Krupke *et al.*¹ make use of the different dielectric constants of metallic and semiconducting nanotubes. In an aqueous suspension, the two types respond to an alternating electric field by moving in opposite directions along the electric-field gradient. The researchers achieved 80% enrichment in metallic nanotubes, which are normally present in a nanotube sample in a 1:2 ratio.

Michael S. Strano *et al.*² propose a chemical separation based on pH titration. They report that the readiness with which the carbon walls of nanotubes are protonated depends on the electronic band gap. Metallic tubes (which have a band gap of zero) can be protonated at near-neutral pH, but increasingly acidic solutions are needed as the band gap increases (as it does in semiconducting nanotubes). The protonation is reversible at alkaline pH. This should allow continuous 'fractionation' of nanotubes according to their band gap. **Philip Ball**

Cancer

From breast to bone

Cancer Cell **3**, 537–549 (2003)

It is not clear how metastasis — the spread of tumour cells to distant organs — occurs. Some believe that most cells in a primary tumour carry a set of metastasis-causing genetic alterations. But from their study of human breast cancer cells, Yibin Kang *et al.* provide evidence to the contrary.

Most patients with advanced breast cancer develop secondary tumours in their bones. Kang *et al.* found that, after inoculating mice with cells from a human breast cancer line, 30% of the animals also went on to develop bone tumours. From these tumours, the authors pulled out a subpopulation of highly metastatic cells and studied their genetic profile. The cells overexpressed several genes, including those involved in forming new blood vessels and the ability to home in on and invade bone. This suggests that key genes may influence the environment around a tumour in favour of metastasis.

Notably, however, only a fraction of the original breast cancer cells had the genetic predisposition to infiltrate bone.

This work reveals a firm link between bone metastasis and the expression of several genes that have previously only been correlated with this process. **Helen R. Pilcher**

Earth science

High expectations

Geophys. Res. Lett. doi:10.1029/2003GL017708 (2003)

Arrays of sensors such as the Southern California Integrated GPS Network (SCIGN), designed for monitoring earthquake-prone faults in southern California, show promise for studying the ionosphere, according to E. Calais and colleagues.

GPS — Global Positioning System — sensors rely on radio signals sent by satellites: as the signals pass through the ionosphere, electromagnetic disturbances in this region can influence signal timing. Events on the ground or in the electrically neutral atmosphere produce upward-moving gravity waves, which affect electron density in the ionosphere.

Calais *et al.* now show that, given a dense enough array of sensors (there are 250 in the SCIGN), it is possible to overcome the small signal-to-noise ratio of such events and identify the direction and speed of the resulting ionospheric disturbances. They were looking for, but failed to find, a 'fingerprint' left in the data by a particular earthquake. Serendipitously, however, they identified other perturbations which they suggest may possibly have been caused by powerful turbulent activity in the lower atmosphere. It is early days. But the approach could prove to be a valuable one in investigating fine-scale atmospheric processes with previously unachievable spatial and temporal resolution. **Tom Clarke**

Quantum physics

How cold can you go?

Phys. Rev. A **67**, 061401(R) (2003)

Bose–Einstein condensation (BEC), creating an exotic and potentially useful quantum state of matter, can be achieved with an unorthodox cooling method, say D. Ivanov and colleagues.

BEC typically requires temperatures of a few billionths of a kelvin. Such cooling is attained in a gas of trapped atoms by, for example, bouncing laser light off the atoms to slow them down. But this and other cooling methods have limitations (for example, in the types of atoms that can be cooled) that could be avoided by stochastic cooling. Here a laser beam 'finds' energetic atoms in a gas cloud and adjusts their momenta selectively — rather like Maxwell's demon. In effect, the beam measures the

momentum (or energy) distribution in a region of gas and then exerts a feedback to shift the mean momentum to zero.

Implementations of stochastic cooling have been proposed previously, but it wasn't clear how cold it could get. Ivanov *et al.* have calculated the quantum limits on the procedure, and find that it can reach the BEC temperature of an atomic gas, despite the heating caused by quantum fluctuations near the condensation point. **Philip Ball**

Developmental biology

Fish in pinstripe suits

Development **130**, 3447–3457 (2003)

Adult zebrafish have stripes running from head to tail. These are formed by alternating rows of two types of pigment cell — blue melanophores and yellow xanthophores. Florian Maderspacher and Christiane Nüsslein-Volhard now show that the interactions between these cells determine the razor-sharp look of the stripes.

The authors first found that pigment cells do not simply 'colour in' a predefined pattern. Stripes form only if both cell types are present and in close proximity. To look for factors that control the cellular interactions at play, the authors then scrutinized two zebrafish mutants, *obelix* and *leopard* (see picture).

Obelix fish have fewer and wider stripes than normal, and, in severe cases, blurred stripe boundaries. The gene affected in *obelix* mutants appears to act only in melanophores, promoting their aggregation. This is necessary to keep the two cell types spatially separated, so affecting the integrity of boundaries. The *leopard* gene is required by both cell types, and its absence in either transforms stripes into spots. The gene seems to control the interactions between cells of the same type and between those of different types, thereby regulating the shape of the boundary.

The molecules encoded by the *obelix* and *leopard* genes are not known, however, nor are they the only factors involved in stripe formation. **Marie-Thérèse Heemels**



How the zebrafish... From top, a normally striped fish, and *obelix* and *leopard* mutants.

Size and growth modification in clownfish

Sex change is not the only way these fish achieve dominance — they grow into the role.

Conflicts of interest are part and parcel of living in a social group, although these can reduce the fitness of individual members. Here I show that clownfish (*Amphiprion percula*) adjust their size and growth rate according to their position in the group hierarchy, maintaining a well-defined size difference with respect to individuals above them in social rank. This strategy to prevent conflict is a surprising departure from the more usual ploy used by many animals of modifying their behaviour within the group^{1,2}.

In Madang Lagoon, Papua New Guinea, groups of *A. percula* inhabit sea anemones, which protect them from predators (Fig. 1)^{3,4}. Each group is composed of a breeding pair and between zero and four non-breeders^{5,6}. Within each group, there is a size-based dominance hierarchy: the female is largest (rank 1), the male is second largest (rank 2), and the non-breeders get progressively smaller as the hierarchy is descended (ranks 3–6; ref. 5). If the female of a group dies, the male changes sex and becomes the breeding female, while the largest non-breeder becomes the breeding male⁵.

Subordinates benefit from settling in an anemone and queuing for breeding positions^{7,8}, their chances of success being determined by their rank within the group^{8,9}. Dominants, however, gain no benefits from subordinates, which are potential challengers for their position (P.B., unpublished results). This asymmetry generates evolutionary conflict over subordinate group membership, and dominants occasionally



Figure 1 Clownfish modify their size, rather than their behaviour, to ensure harmony in social groups.

evict¹⁰ or kill³ subordinates that are of similar size to themselves. Conflict would be avoided, however, if a size difference could be maintained between individuals adjacent in rank that meant that a subordinate would never present a threat to its immediate dominant.

A multivariate analysis reveals that there are well-defined size differences between individuals adjacent in rank (Fig. 2a): the standard length of subordinates (ranks 2–6) is positively related to their rank, the standard length of rank 1, and the number of individuals in the group (SAS mixed

procedure: rank, $F_{4,137} = 325.71$, $P < 0.0001$; standard length rank 1, $F_{1,137} = 105.37$, $P < 0.0001$; number of individuals, $F_{4,137} = 14.13$, $P < 0.0001$; $R^2 = 0.82$), whereas anemone diameter, depth and reef had no significant effect.

These well-defined size differences are maintained by the regulation of growth in response to social context, as illustrated by an experiment in which the rank-2 fish is removed from the group (Fig. 2b). In this experiment, the growth of rank-3 fish is positively related to the removal of the fish of rank 2 and the growth of the fish of rank 1, and negatively related to its own residual standard length (the difference between observed and expected measures) (SAS mixed procedure: removal, $F_{1,48} = 255.82$, $P < 0.0001$; growth of rank 1, $F_{1,48} = 6.22$, $P = 0.0161$; residual standard length, $F_{1,48} = 76.12$, $P < 0.0001$; $R^2 = 0.89$), whereas its initial standard length, anemone diameter, depth and reef had no significant effect.

The removal of the rank-2 fish caused the rank-3 fish to ascend in rank. Subordinates that ascended in rank grew more than individuals that did not ascend in rank. The residual standard length indicates how much smaller (negative residuals) or larger (positive residuals) the subordinate's observed standard length was compared with its standard length predicted by its social context (Fig. 2a). When the initial standard length of a subordinate was less than predicted, it grew more than an individual whose initial standard length was greater than predicted.

The coefficients of the size analysis

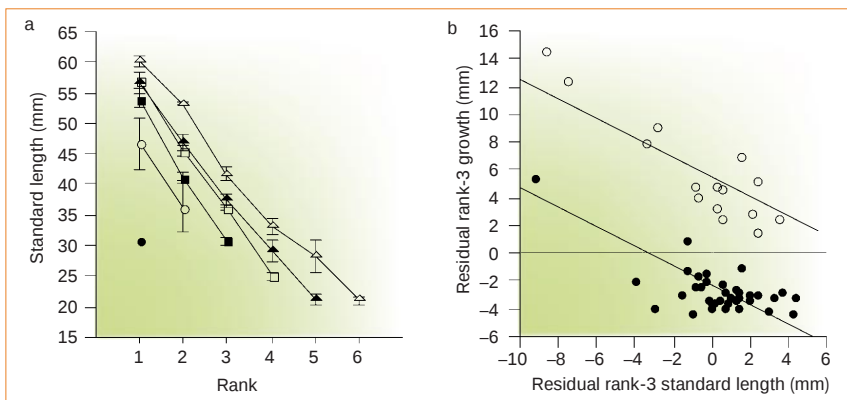


Figure 2 Size and growth of clownfish (*Amphiprion percula*) in relation to rank. **a**, Standard length (SL) of subordinates (ranks 2–6) is related to their rank, the SL of rank 1, and the number of individuals in the group ($n = 237$ individuals in 96 groups). Symbols represent the mean SL for each rank and group size (different symbols correspond to different group sizes of one to six fish); bars represent standard error. SL was measured in January 1997. **b**, Growth of fish of rank 3 was related to the removal of the fish of rank 2, the growth of the rank-1 fish, and their own 'residual' SL (control, filled circles, $n = 36$; experimental, open circles, $n = 16$): residual rank-3 SLs are the residuals of the model demonstrated in **a**. Residual rank-3 growth values are the residuals of a reduced model showing that the growth of rank 3 is related to the growth of rank 1. Fish of rank 2 were removed in February 1997; growth was calculated as the change in SL between January 1997 and December 1997. Individuals were identified by the natural variation in their colour markings.

indicate that the average difference in standard length between fish of ranks 2 and 3 was 10.1 mm (Fig. 2a). The coefficients of the growth analysis indicate that individuals of rank 3 that were experimentally elevated to rank 2 had the same growth rate as control individuals (rank 3) when their standard length was 10.8 mm greater than that of controls (Fig. 2b). The similarity of these two estimates shows that the size difference that is found between individuals adjacent in rank would be restored after the disappearance of intermediate individuals.

These findings show that there are well-defined size differences between individuals adjacent in rank within groups of *A. percula*, and that these are maintained by the precise regulation of subordinate growth. The maintenance of size differences may resolve evolutionary conflict over group membership, because subordinates do not become a threat to their dominants. The results indicate that the growth rate

and the size adopted by any group-living organism could be a strategic response to its social environment.

Peter Buston

Department of Neurobiology and Behavior,
Cornell University, Seeley G. Mudd Hall, Ithaca,
New York 14853, USA

Present address: National Center for Ecological
Analysis and Synthesis, University of California,
Santa Barbara, California 93101, USA

e-mail: buston@nceas.ucsb.edu

1. Reeve, H. K. *Nature* **358**, 147–149 (1992).
2. Balshine-Earn, S., Neat, F. C., Reid, H. & Taborsky, M. *Behav. Ecol.* **9**, 432–438 (1998).
3. Allen, G. R. *The Anemonefishes: Their Classification and Biology* (T. F. H. Publications, Neptune City, New Jersey, 1972).
4. Elliott, J. K. & Mariscal, R. N. *Mar. Biol.* **138**, 23–36 (2001).
5. Fricke, H. W. Z. *Tierpsychol.* **50**, 313–326 (1979).
6. Fautin, D. G. *Symbiosis* **14**, 143–160 (1992).
7. Kokko, H. & Johnstone, R. A. *Proc. R. Soc. Lond. B* **266**, 571–578 (1999).
8. Buston, P. M. *Proc. R. Soc. Lond. B* (submitted).
9. Field, J., Shreeves, G. & Sumner, S. *Behav. Ecol. Sociobiol.* **45**, 378–385 (1999).
10. Buston, P. M. *Behav. Ecol.* (in the press).

Competing financial interests: declared none.

Hydrogen bonding

Single enantiomers from a chiral-alcohol catalyst

Hydrogen bonding acts as a ubiquitous glue to sustain the intricate architecture and functionality of proteins, nucleic acids and many supramolecular assemblies^{1,2}, but this weak interaction is seldom used as a force for promoting chemical reactions^{3–5}. Here we show that a simple chiral alcohol uses hydrogen bonding to catalyse an important family of cycloaddition reactions of a diene with various aldehydes — moreover, this reaction is highly enantioselective, generating only one of the mirror-image forms of each dihydropyran product. This type of catalysis mimics the action of enzymes and antibodies, and is unlike traditional, metal-based catalysts used in organic chemistry⁶.

Important biological molecules such as DNA and proteins, and therefore many pharmaceutical drugs, are chiral — that is,

they are not superimposable on their mirror image (the pair of asymmetric molecules are known as enantiomers), so any chemical reaction that can selectively synthesize one enantiomeric form of a chiral compound is potentially very useful.

Candidate catalysts for such reactions are generally based on Lewis-acid metals⁶. On the basis of our discovery that hetero-Diels–Alder (HDA) reactions⁷ between unactivated ketones and 1-amino-3-siloxy diene (compound **1**) are accelerated in protic solvents, we investigated the use of chiral alcohols for the asymmetric catalysis of these cycloadditions⁸. Of the possible alcohols that could be used for the HDA reaction between diene **1** and benzaldehyde, the TADDOL ($\alpha,\alpha,\alpha',\alpha'$ -tetraaryl-1,3-dioxolan-4,5-dimethanol) class of chiral alcohols, which are known to complex with carbonyl groups, was considered promising⁹.

Figure 1 shows the scheme of the HDA reaction in the presence of this chiral alcohol. A solution of 0.1 mmol (*R,R*)-1-naphthyl TADDOL (**2**) and benzaldehyde

(1.0 mmol) in toluene at -78°C was treated with diene **1** (0.5 mmol) and stirred, causing a smooth HDA reaction to take place. Analysis of the reaction mixture by ^1H NMR indicated that the cycloadduct had been formed as a single diastereomer (**3**, where R is C_6H_5), tentatively assigned as *endo*. On treatment with acetyl chloride (1.0 mmol), the cycloadduct was converted to dihydropyrene (**4**, 70% overall yield).

Analysis of the product by high-performance liquid chromatography revealed that the *S*-enantiomer had been produced preferentially over the *R*-enantiomer (>99:1). The reaction is considerably accelerated by TADDOL (**2**): in its absence, there was no reaction under otherwise identical conditions. Moreover, the monomethyl and dimethylether derivatives of **2** were poor catalysts, indicating that the hydrogen-bonding capability of **2** is crucial for the catalytic function.

This metal-free asymmetric catalysis, which does not involve a covalent connection between the catalyst and the reactant¹⁰, can be used for cycloadditions between **1** and a range of aldehydes with different R-groups (Fig. 1). Aromatic aldehydes were particularly effective as dienophiles in these HDA reactions. The resultant dihydropyrene products (see **4a–f** in supplementary information) were also consistently obtained with high enantiomer ratios. Aldehydes with aliphatic (**4g**) and α,β -unsaturated (**4h**) R-groups could be used successfully in these reactions.

Our results show that hydrogen bonding by a simple chiral alcohol to a carbonyl group can accomplish what has previously been considered to be in the domain of enzymes, catalytic antibodies and chiral metal-based Lewis acids. These studies indicate the broad potential for hydrogen-bond catalysis in asymmetric synthesis.

**Yong Huang, Aditya K. Unni,
Avinash N. Thadani, Vireh H. Rawal**

Department of Chemistry, University of Chicago,
Chicago, Illinois 60637, USA

e-mail: vrawal@uchicago.edu

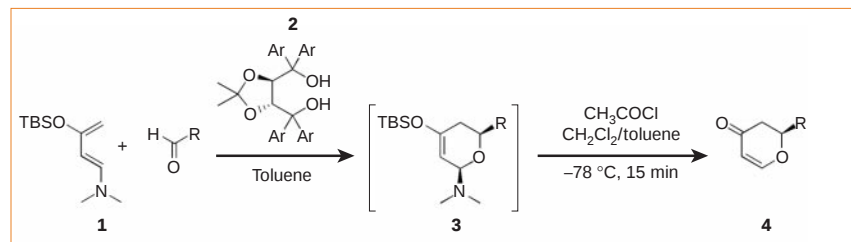


Figure 1 Asymmetric hetero-Diels–Alder reaction of diene **1** with aldehydes to produce dihydropyrenes (**4**). The reaction is catalysed by hydrogen bonding and yields specific enantiomeric molecules (see supplementary information for eight examples of products — molecules **4a–h** — bearing different R-groups). TBS, *tert*-butyldimethylsilyl derivative; Ar, 1-naphthyl group. The experimental procedure involved adding diene (**1**, 0.5 mmol) to a solution of TADDOL (**2**, 0.1 mmol) and the aldehyde (1.0 mmol) in toluene (0.5 ml) cooled to -40°C (-78°C for **4a**, **4c** and **4f**). The mixture was stirred for 24 h at -40°C (48 h for **4a**, **4c** and **4f**) and then diluted with CH_2Cl_2 (2.0 ml). Acetyl chloride (1.0 mmol) was added dropwise at -78°C , and the mixture was stirred for a further 15 min and then separated by chromatography on silica gel (yields 52–97%; enantiomer ratio 96:4–99:1).

1. Prins, L. J., Reinhoudt, D. N. & Timmerman, P. *Angew. Chem. Int. Edn* **40**, 2383–2426 (2001).
2. Martin, T. W. & Derewenda, Z. S. *Nature Struct. Biol.* **6**, 403–406 (1999).
3. Kelly, T. R., Meghani, P. & Ekkundi, V. S. *Tetrahedron Lett.* **31**, 3381–3384 (1990).
4. Vachal, P. & Jacobsen, E. N. *J. Am. Chem. Soc.* **124**, 10012–10014 (2002).
5. Schuster, T., Bauch, M., Dürner, G. & Göbel, M. W. *Org. Lett.* **2**, 179–181 (2000).
6. Jacobsen, E. N., Pfaltz, A. & Yamamoto, H. (eds) *Comprehensive Asymmetric Catalysis Vols 1–3* (Springer, Berlin, 1999).
7. Jørgensen, K. A. *Angew. Chem. Int. Edn* **39**, 3558–3588 (2000).
8. Huang, Y. & Rawal, V. H. *J. Am. Chem. Soc.* **124**, 9662–9663 (2002).
9. Seebach, D., Beck, A. K. & Heckel, A. *Angew. Chem. Int. Edn* **40**, 92–138 (2001).
10. Dalko, P. I. & Moisan, L. *Angew. Chem. Int. Edn* **40**, 3726–3748 (2001).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

Transcription regulation and animal diversity

Michael Levine* & Robert Tjian*†

* Department of Molecular and Cell Biology, Division of Genetics and Development, Center for Integrative Genomics, and

† The Howard Hughes Medical Institute, University of California, Berkeley, 401 Barker Hall, Berkeley, California 94720, USA

Whole-genome sequence assemblies are now available for seven different animals, including nematode worms, mice and humans. Comparative genome analyses reveal a surprising constancy in genetic content: vertebrate genomes have only about twice the number of genes that invertebrate genomes have, and the increase is primarily due to the duplication of existing genes rather than the invention of new ones. How, then, has evolutionary diversity arisen? Emerging evidence suggests that organismal complexity arises from progressively more elaborate regulation of gene expression.

Comparative genome analyses indicate that increases in gene number do not account for increases in morphological and behavioural complexity. For example, the simple nematode worm, *Caenorhabditis elegans*, possesses nearly 20,000 genes¹ but lacks the full range of cell types and tissues seen in the fruitfly *Drosophila*, which contains fewer than 14,000 genes². Indeed, the revelation that the human genome contains only ~30,000 protein-coding genes precipitated a frenzy of speculation regarding the molecular basis of organismal complexity^{3,4}.

In principle, there are many ways in which a relatively small number of genes could be exploited so as to generate more complex organisms over evolutionary time. Two mechanisms which have been posited as important sources of complexity are alternative splicing—the production of different RNA species from a given gene during mRNA splicing—and DNA rearrangement, where genes themselves are rearranged during cellular differentiation, as used to generate diversity in mammalian immune systems^{5–6}. However, we argue here that physiological and behavioural complexity correlates with the likely number of gene expression patterns exhibited during an animal's life cycle. We discuss two pervasive mechanisms for producing complexity at the genetic level—greater elaboration of *cis*-regulatory DNA sequences, which control the expression of nearby genes, and increased complexity in the multi-protein transcription complexes that regulate gene expression.

An expansion in regulatory complexity

Protein coding sequences account for only a small fraction of a typical metazoan genome; less than 2% in the case of the human genome. It is difficult to estimate the *cis*-regulatory content of metazoan genomes, but it is easy to imagine that as much as a third of the human genome, a remarkable one billion base pairs, controls chromosome replication, condensation, pairing, and segregation, and most importantly for our present discussion, gene expression. Even simple creatures like the sea squirt, *Ciona intestinalis*, have an estimated 10,000–20,000 tissue-specific enhancers⁷. A typical genetic locus in *Drosophila* contains several enhancers (along with other *cis*-regulatory DNAs such as insulators) scattered over an average distance of 10 kilobases (kb) of genomic DNA⁸ with transcribed DNA comprising just 2 or 3 kb. By contrast, the regulatory DNAs of unicellular eukaryotes such as yeast are usually composed of short sequences (a few hundred base pairs, bp) located immediately adjacent to the core promoter⁹.

Commensurate with this increased complexity in *cis*-control elements, 5–10% of the total coding capacity of metazoans is dedicated to proteins that regulate transcription. These proteins fall into several major classes: First there are the sequence-specific DNA binding proteins that mediate gene-selective transcriptional activation or repression. Second, there are the general but diverse components of large multi-protein RNA polymerase machines

required for promoter recognition and the catalysis of RNA synthesis. Finally, there are the chromatin remodelling and modification complexes that assist the transcriptional apparatus to navigate through chromatin.

The yeast genome encodes a total of ~300 transcription factors; this includes both sequence-specific DNA-binding proteins and subunits of general transcription complexes such as TFIID⁹. In contrast, the genome sequences of *C. elegans* and *Drosophila* reveal at least 1,000 transcription factors in each organism^{1,10}. There may be as many as 3,000 transcription factors in humans⁴. It would appear that organismal complexity correlates with an increase in both the ratio and absolute number of transcription factors per genome. Yeast contains an average of one transcription factor per 20 genes, while humans appear to contain one factor for every ten genes. Given the combinatorial nature of transcription regulation, even this twofold increase in the number of factors could produce a dramatic expansion in regulatory complexity, as we discuss below.

Regulatory DNA sequences

A typical yeast transcription unit is presented in Fig. 1a. The regulation of gene expression usually depends on DNA sequences located immediately 5' of the transcription start site (labelled '+1'). Most core promoters contain a TATA element, which serves

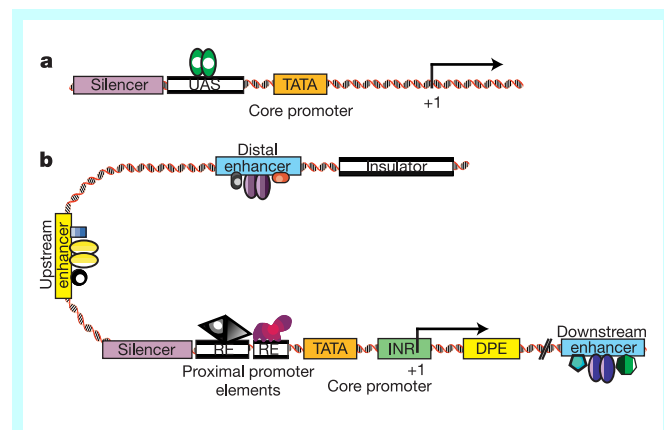


Figure 1 Comparison of a simple eukaryotic promoter and extensively diversified metazoan regulatory modules. **a**, Simple eukaryotic transcriptional unit. A simple core promoter (TATA), upstream activator sequence (UAS) and silencer element spaced within 100–200 bp of the TATA box that is typically found in unicellular eukaryotes. **b**, Complex metazoan transcriptional control modules. A complex arrangement of multiple clustered enhancer modules interspersed with silencer and insulator elements which can be located 10–50 kb either upstream or downstream of a composite core promoter containing TATA box (TATA), Initiator sequences (INR), and downstream promoter elements (DPE).

as a binding site for TBP (TATA-binding protein)¹¹. In general, promoters are selected for expression by the binding of TBP to the TATA element. The regulation of TBP binding depends on upstream activating sequences (UAS), which are usually composed of 2 or 3 closely linked binding sites for one or two different sequence-specific transcription factors¹². A few genes in the yeast genome, such as *HO*, contain distal regulatory sequences¹³, but the vast majority contains a single UAS located within a few hundred base pairs of the TATA element.

Metazoan genes contain highly structured regulatory DNAs that direct complex patterns of expression in many different cell types during development (Fig. 1b)⁸. A typical animal gene is likely to contain several enhancers that can be located in 5' and 3' regulatory regions, as well as within introns. Each enhancer is responsible for a subset of the total gene expression pattern; they usually mediate expression within a specific tissue or cell type. A typical enhancer is something like 500 bp in length and contains on the order of ten binding sites for at least three different sequence-specific transcription factors, most often two different activators and one repressor⁸. The core promoter is compact and composed of ~60 bp straddling the transcription start site. There are at least three different sequence

elements that can recruit the TBP containing TFIID initiation complex: TATA, initiator element (INR) and the downstream promoter element (DPE)¹⁴. Many genes contain binding sites for proximal regulatory factors located just 5' of the core promoter. These factors do not always function as classical activators or repressors; instead, they might serve as 'tethering elements' that recruit distal enhancers to the core promoter^{15,16}. Finally, insulator DNAs prevent enhancers associated with one gene from inappropriately regulating neighboring genes¹⁷. These regulatory DNAs, enhancers, silencers and insulators are scattered over distances of roughly 10 kb in fruitflies and 100 kb in mammals.

This elaborate organization of the regulatory DNA permits the detailed control of gene expression. Indeed, a defining feature of metazoan gene regulation is the use of multiple enhancers, silencers and promoters to control the activities of a single transcription unit. Enhancers were initially identified and characterized in mammalian viruses and cultured cells¹⁸. They were shown to be composed of multiple binding sites for different regulatory proteins. Subsequent analyses in transgenic animals reveal an even greater level of complexity⁸. Metazoan enhancers integrate different regulatory inputs, such as those produced by multiple signalling pathways, to direct stripes, bands, and tissue-specific patterns of gene expression in *Drosophila* embryos and imaginal disks. Similarly organized enhancers are responsible for the restricted expression of the mouse *Pit-1* gene in the anterior pituitary¹⁹, the localized expression of *Hox* genes in rhombomeres of the vertebrate hind-brain²⁰ and the selective expression of immunoglobulin genes in mammalian B lymphocytes²¹.

Tissue-specific enhancers can work over distances of 100 kb or more. There are numerous examples of long-range gene regulation in flies and mammals. For example, the embryonic enhancers that regulate the mouse and human *Igf-2* gene map over 100 kb from the transcription start site^{22,23}. The wing margin enhancer that regulates the *Drosophila cut* gene maps at a similar distance²⁴, while the tissue-specific enhancers that regulate the *Drosophila Decapentaplegic* gene (*Dpp*), and orthologous genes in vertebrates, map far downstream of the transcription unit^{25,26}. This type of long-range regulation is not observed in yeast and might be a common feature of genes that play critical roles in morphogenesis and are therefore subject to stringent regulation.

Enhancers generate complex gene expression patterns

Different enhancers can work independently of one another to direct composite patterns of gene expression when linked within a common *cis*-regulatory region. For example, the seven stripes of *even-skipped* expression in the *Drosophila* embryo depend on five separate enhancers; two located 5' of the transcription start site and three located 3' of the gene^{27,28}. These enhancers function in an autonomous fashion owing to short-range repression: sequence-specific repressors bound to one enhancer do not interfere with the activities of neighbouring enhancers. The repressors must bind within 50–100 bp of an upstream activator or the core promoter in order to inhibit expression²⁹.

Additional diversity in gene regulation is achieved by the use of multiple promoters for a single gene. For example, the segmentation gene *Hunchback* contains a maternal promoter that is ubiquitously expressed in the germline and a separate zygotic promoter, which mediates restricted expression in the anterior half of early embryos³⁰.

A potential 'trafficking' problem arises from the fact that *cis*-regulatory DNAs can map far from their target promoters. In some cases the regulatory DNAs actually map closer to the 'wrong' promoter than the proper one²⁵. There are at least three underlying mechanisms for ensuring that the right enhancer interacts with the right promoter: insulator DNAs, gene competition, and promoter-proximal tethering elements that recruit distal enhancers.

Insulators are typically 300 bp to 2 kb in length and often contain

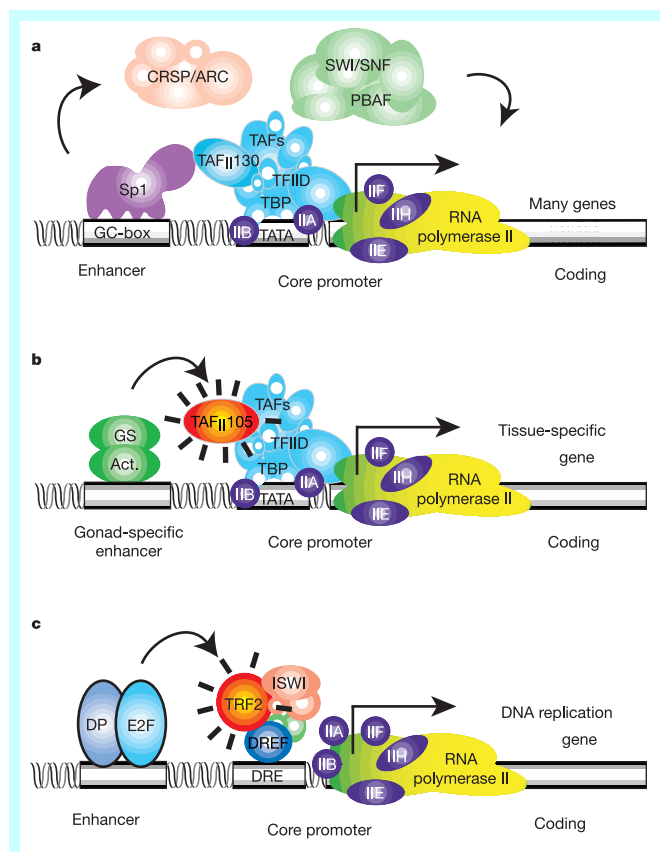


Figure 2 The multi-subunit general transcription apparatus: identification of tissue-specific and gene-selective subunits. Diversified metazoan transcription initiation complexes. **a**, The eukaryotic transcriptional apparatus can be subdivided into three broad classes of multi-subunit ensembles that include the RNA polymerase II core complex and associated general transcription factors (TFIIA, -B, -D, -E, -F and -H), multi-subunit cofactors (mediator, CRSP, TRAP, ARC/DRIP, and so on) and various chromatin modifying or remodelling complexes (SWI/SNF, PBAF, ACF, NURF and RSF). **b, c**, Metazoan organisms have evolved multiple gene-selective and tissue-specific TFIID-like assemblies by using alternative TAFs (TBP-associated factors such as the ovarian-specific TAF105) as well as TRFs (TBP-related factors such as TRF2 in *Drosophila* and mice) to mediate the formation of specialized RNA polymerase initiation complexes that direct the transcription of tissue-specific and gene-selective programmes of expression.

clustered binding sites for large zinc finger proteins, such as Su(Hw) and CTCF^{17,31}. They selectively block the long-range interaction of a distal enhancer with a proximal target promoter when positioned between the two. Insulators were first identified at gene boundaries, where they prevent *cis*-regulatory sequences in one gene from inappropriately interacting with neighbouring loci³². Insulators have also been identified within complex genetic loci, including the *Bithorax* complex in *Drosophila* and the *Igf-2* locus in mice³¹.

Gene competition was first documented in the chicken globin locus, and occurs when a shared enhancer preferentially interacts with just one of two linked promoters³³. There is emerging evidence that selectivity depends on *cis*-elements within the core promoter, particularly the TATA sequence, INR and DPE¹⁴ (Fig. 1). Core promoters that lack a TATA sequence usually contain a compensatory DPE element, in order to ensure recognition by the RNA

polymerase II (Pol II) transcription complex. It is possible that some enhancers preferentially activate TATA-containing promoters, while others activate DPE-containing promoters^{34,35}. The binding of Pol II depends on an associated multi-subunit complex, TFIID, which is composed of TBP, and TBP-associated factors (TAFs) (Fig. 2A). TFIID binds TATA via TBP, and interacts with DPE and INR through different TAFs¹⁴. It seems likely that different core promoters can interact with distinct TFIID-like initiator complexes, thus providing a mechanism for matching subsets of enhancers to specific core promoters at complex transcription units (Fig. 2; see below).

There is also evidence that the selection of a particular target promoter by a shared enhancer depends on regulatory factors that bind in promoter-proximal regions, just 5' of the core promoter³⁶. It is conceivable that these proximal proteins do not function as classical transcriptional activators or repressors. Instead, they might serve as tethering elements that recruit specific distal enhancers^{15,16}. It is easy to imagine a promoter-proximal 'code', whereby one combination of regulatory factors recruits some enhancers, while another combination recruits other enhancers.

In summary, the *cis*-regulatory DNA of higher metazoans is highly structured and exhibits a modular organization consisting of insulators, silencers, enhancers, and discrete core promoters. We now consider the different protein complexes that interact with these regulatory DNAs. There are three major strategies for regulating the binding and function of the RNA Pol II complex at the core promoter. First, divergent TFIID complexes bind specific sequence elements within the core promoter and recruit Pol II. Second, multi-subunit transcription complexes that are related to the yeast mediator complex also facilitate the binding and function of Pol II. Third, there are enzymatic complexes that remodel or modify (for example, acetylation) chromatin. We discuss each of these pathways of gene activation, with a particular emphasis on diverse TFIID and multi-subunit transcription complexes.

Diversification of core promoter transcription complexes

Multi-subunit transcription complexes are recruited to distal enhancers through interactions with sequence-specific transcriptional activators. These complexes facilitate the binding and function of Pol II at the core promoter (Fig. 2a). Initial studies on the Pol II complex suggested close conservation of the core RNA polymerase, as well as the accessory factors required for promoter recognition and transcription, including TFIIA, -B, -D, -E, -F, and -H (ref. 37). However, it is now apparent that metazoans have evolved multiple related TFIID complexes that can function at distinct promoters through the use of tissue-specific TAFs and TRFs which are not found in yeast.

A tissue-specific mammalian TAF, TAF_{II}105, which is related to the ubiquitously expressed human TAF_{II}130, operates as part of a unique TFIID complex in follicle cells of the ovary, thereby permitting the selective activation of a small subset of genes³⁸ (Fig. 2b). Similarly, a testes-specific homologue of TAF_{II}80 (Cannonball) is required for spermatogenesis in *Drosophila*³⁹, and there is emerging evidence that Cannonball may be a component of a complete sperm-specific TFIID complex.

Diverse TFIID complexes have also evolved through the duplication of TBP. There is only one TBP gene and no TRFs in yeast, but there are four TRFs (TRF1–4) in addition to TBP in *Drosophila*^{40–46}. TRF1 binds poorly to TATA sequences, but instead initiates transcription at a small group of TATA-less Pol II core promoters. TRF1 is known only in insects, whereas orthologues of TRF2 have also been identified in *C. elegans*, frogs, mice and humans^{41–43} (Fig. 2). TRF2 might be essential for mammalian spermiogenesis^{44,45} but exerts a more pervasive influence on gene expression during the early embryonic development of *C. elegans*, *Drosophila*, zebrafish, and *Xenopus*^{42,43,46}. *Drosophila* TRF2 does not bind TATA but is targeted to distinct core promoters by its interaction with a partner,

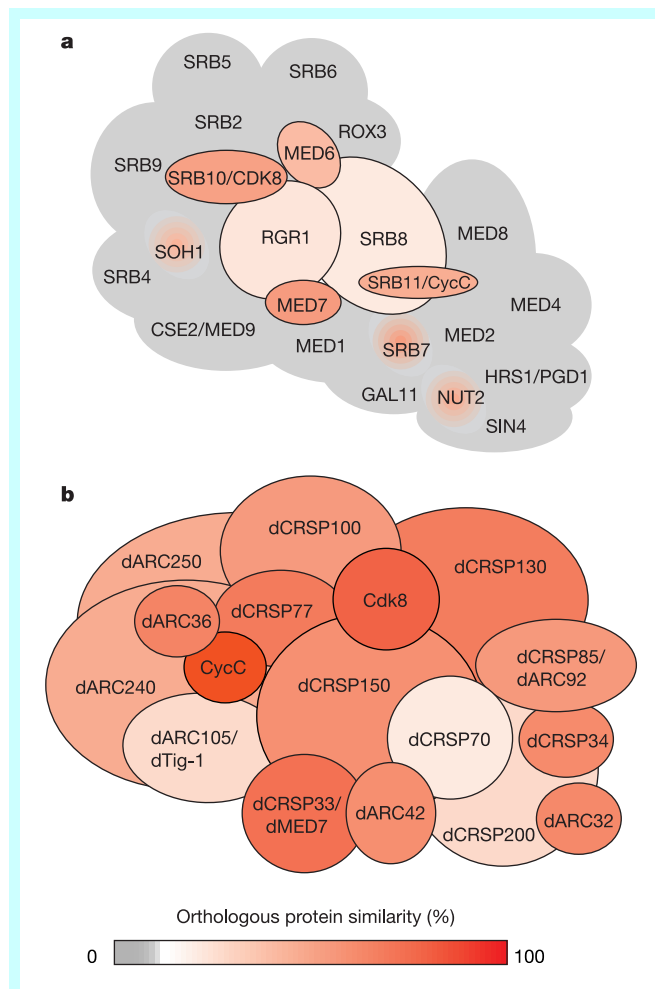


Figure 3 Diversification of cofactor complexes. Critical components of large multi-subunit cofactor complexes such as mediator and CRSP/ARC have diversified both structurally and functionally, particularly when comparing uni-cellular eukaryotes to their metazoan counterparts. **a**, Yeast mediator versus human cofactor complex. Most of the yeast mediator subunits display very limited, if any, sequence similarity to subunits of human CRSP/ARC although these two cofactor complexes are thought to function in an analogous manner to potentiate RNA Pol II transcription. **b**, *Drosophila* versus human cofactor complex. Even some of the CRSP/ARC subunits have significantly diverged between *Drosophila* and human, although many share conserved regions. We have colour-coded the extent of amino acid sequence similarity between different orthologous co-activator subunits from yeast, *Drosophila* and human. Grey is used to depict yeast subunits with no counterparts in *Drosophila* or humans. White or lightly shaded components represent very little structural conservation (17–20%) while darkly shaded orange and red represent highly conserved (40–80%) subunits.

DREF⁴⁶, which selectively activates genes encoding DNA replication proteins (Fig. 2c). These findings suggest that metazoans have evolved TFIID-related transcription complexes responsible for recognizing distinct core promoters with specific regulatory activities. As mentioned earlier, a number of these genes contain multiple tandem promoters^{40,46}. These may be recognized by distinct TFIID-related complexes in different cell types.

Transcription cofactor complexes

The yeast mediator is a multi-subunit co-activator complex that is thought to facilitate the binding and/or function of Pol II at the core promoter^{47,48}. While yeast has one such cofactor complex, metazoans contain several related complexes: TRAP, CRSP, ARC/DRIP, SMCC and hMed. These complexes are recruited to the DNA template via interactions with a variety of sequence-specific transcriptional activators, including nuclear receptors such as the vitamin D receptor and the thyroid hormone receptor^{49–52}. Like TFIID, these cofactor complexes might serve as bridges between distal activators and the core promoter. However, they might not function solely through the simple recruitment of Pol II, but can also be induced to undergo conformational changes that may be essential for activating transcription⁵³. Some of these cofactor complexes could also function in transcriptional repression⁵². In the case of human ARC and CRSP, the smaller CRSP complex can augment transcription *in vitro*, while the larger ARC complex might be involved in repression⁵³.

Protein chromatography and functional transcription assays are identifying a rapidly expanding number of metazoan cofactor complexes. These represent some of the most dramatic examples of the diversification of general transcription complexes in evolution^{53–56}. In contrast to other components of the general transcription machinery such as RNA Pol II subunits, TFIIB, -D, -E, -F and -H, most of the protein subunits that comprise the yeast mediator and metazoan cofactor complexes are not highly conserved (Fig. 3a). In fact, only two of the subunits show significant structural conservation, and these correspond to the smallest subunits (yeast Med6/Med7 and human CRSP33/34)^{55–57}. The remaining subunits share almost no discernible sequence similarity save for one or two short stretches of 20–30 amino acid residues (for example, only 10% of the 1,507 residues are conserved between human CRSP 150 and yeast RGR1). Moreover, the low-resolution structure of the yeast mediator complex does not obviously resemble the corresponding human complexes^{53,57}. Thus, in contrast to the strong conservation of TFIID and Pol II subunits, cofactor complexes have diversified extensively between eukaryotes, and have expanded among metazoans^{58,59}.

Diversification of cofactor complexes might also reflect the increasing variety of sequence-specific activators seen in different metazoans. A comparison of the fruitfly and human ARC/CRSP subunits reveals roughly 50% overall similarity in amino acid sequence (Fig. 3b). However, close comparisons of some of the orthologues reveal interesting differences. For example, the *Drosophila* ARC 70, -130 and -230 subunits contain an expansion of glutamine residues, a prevalent feature of sequence-specific activators in *Drosophila*, whereas the human CRSP 70, -130 and -230 subunits possess a broad spectrum of co-activation sequence motifs^{55,56}.

Even a twofold increase in the number of potential cofactors would result in a substantial increase in the combinatorial control of gene expression. Concomitant with this diversification of cofactor subunits is an increase in the size of gene families encoding sequence-specific transcription factors. Single genes in *Drosophila* are often related to an expanded family of similar genes in humans⁶⁰. Different family members come to acquire diverse activation domains, while retaining highly conserved DNA binding domains. This increase in the complexity of activation domains parallels the expansion and diversification of cofactor subunits.

Chromatin remodelling and modifying complexes

The enzyme complexes that either displace (remodel) nucleosomes (for example, Swi/Snf, Baf/Brm, Acf and Nurf) or covalently modify histones via acetylation (for example, the histone acetyl transferases CBP/p300/pCAF, GCN5) represent another potential source of regulatory diversification during metazoan evolution^{61–65}. Although such complexes are found in yeast, there is only limited conservation of orthologous subunits in mammals. For instance, there are a number of *Drosophila* and human ISWI-containing chromatin remodelling complexes such as NURF and RSF^{66,67} with no apparent yeast homologues. There is also emerging evidence that remodelling complexes such as the BAFs have diversified in mammals along with the acquisition of specialized cell types. For example, a neuron-specific BAF complex in mammals has no obvious counterpart in fruitflies or nematodes⁶⁸. Moreover, some of the remodelling complexes mediate transcriptional repression rather than activation. For example, the MBD2 protein mediates repression on methylated DNA templates by recruiting NuRD⁶⁵, a complex that contains both a Swi/Snf nucleosome displacement activity and histone deacetylase (HDAC) activity. Similarly, in *Drosophila*, the Mi-2 complex initiates the transcriptional silencing of homeotic genes by Polycomb; it contains an ATP-dependent nucleosome displacement factor, as well as an HDAC⁶⁹.

Perspective

Rapidly evolving cofactor and chromatin remodelling complexes might be critical for the integration of complex *cis*-regulatory information in metazoans, such as long-range enhancer–promoter interactions. Altogether, the expanded number of metazoan TFIID, cofactor, and chromatin remodelling/modifying complexes appears to be commensurate with the diversification of sequence-specific activators, including nuclear hormone receptors, STATS, SMADs, and promoter-proximal factors such as Sp1, CTF and NTF, all of which are unique to metazoans. This diversification is probably essential for implementing regulation by the highly sophisticated *cis*-regulatory DNAs seen in higher metazoans.

It is unlikely that any single core promoter requires all the transcription complexes we have described. Instead, we suggest that different complexes might be required for regulating distinct *cis*-DNA elements in a temporal and tissue-specific manner. For example, different core promoters might be recognized by diverse TFIID complexes, while enhancers might interact with different cofactor complexes. This type of mixing and matching could generate a huge repertoire of distinct gene expression patterns.

Assembled metazoan genome sequences are now available for nematode worms, fruitflies, mosquitoes, sea squirts, pufferfish, mice and humans. These assemblies provide a powerful foundation for the comparative analysis of gene regulation networks. As we enter the post-genome era it is possible to envision the elucidation of a *cis*-regulatory code, whereby different classes of *cis*-DNAs can be identified by simple sequence analysis. Indeed, computational methods have been used to identify novel enhancers in the *Drosophila* genome based on the clustering of *cis*-regulatory elements^{70–74}. Increasingly more powerful methods of comparative genomics should identify many of the changes in *cis*-regulatory DNAs and general transcription complexes underlying animal diversity. □

doi:10.1038/nature01763.

1. Ruvkun, G. & Hobert, O. The taxonomy of developmental control in *Caenorhabditis elegans*. *Science* **282**, 2033–2041 (1998).
2. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
3. Baltimore, D. Our genome unveiled. *Nature* **409**, 814–816 (2001).
4. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
5. Graveley, B. R. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* **17**, 100–107 (2001).
6. Agrawal, A., Eastman, Q. M. & Schatz, D. G. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature* **394**, 744–751 (1998).
7. Harafuji, N., Keys, D. N. & Levine, M. Genome-wide analysis of tissue-specific enhancers in the *Ciona* tadpole. *Proc. Natl Acad. Sci.* **99**, 6802–6805 (2002).

8. Davidson, E. H. *Genomic Regulatory Systems: Development and Evolution* (Academic, New York, 2001).
9. Wyrick, J. J. & Young, R. A. Deciphering gene expression regulatory networks. *Curr. Opin. Genet. Dev.* **12**, 130–136 (2002).
10. Aoyagi, N. & Wassarman, D. A. Genes encoding *Drosophila melanogaster* RNA polymerase II general transcription factors: diversity in TFIIA and TFIID components contributes to gene-specific transcription regulation. *J. Cell Biol.* **150**, F45–F50 (2000).
11. Struhl, K., Kadosh, D., Keaveney, M., Kuras, L. & Moqtaderi, Z. Activation and repression mechanisms in yeast. *Cold Spring Harb. Symp. Quant Biol.* **63**, 413–421 (1998).
12. de Bruin, D., Zaman, Z., Liberatore, R. A. & Ptashne, M. Telomere looping permits gene activation by a downstream UAS in yeast. *Nature* **409**, 109–113 (2001).
13. Brand, A. H., Breeden, L., Abraham, J., Sternglanz, R. & Nasmyth, K. Characterization of a “silencer” in yeast: a DNA sequence with properties opposite to those of a transcriptional enhancer. *Cell* **41**, 41–48 (1985).
14. Smale, S. T. & Kadonaga, J. T. The RNA polymerase II core promoter. *Annu. Rev. Biochem.* (in the press).
15. Su, W., Jackson, S., Tjian, R. & Echols, H. DNA looping between sites for transcriptional activation: self-association of DNA-bound Sp1. *Genes Dev.* **5**, 820–826 (1991).
16. Calhoun, V. C., Stathopoulos, A. & Levine, M. Promoter-proximal tethering elements regulate enhancer-promoter specificity in the *Drosophila* Antennapedia complex. *Proc. Natl Acad. Sci. USA* **99**, 9243–9247 (2002).
17. Burgess-Beusse, B. *et al.* The insulation of genes from external enhancers and silencing chromatin. *Proc. Natl Acad. Sci. USA* **99**, 16433–16437 (2002).
18. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981).
19. DiMattia, G. E. *et al.* The Pit-1 gene is regulated by distinct early and late pituitary-specific enhancers. *Dev. Biol.* **182**, 180–190 (1997).
20. Vesque, C. *et al.* Hoxb-2 transcriptional activation in rhombomeres 3 and 5 requires an evolutionarily conserved cis-acting element in addition to the Krox-20 binding site. *EMBO J.* **15**, 5383–5396 (1996).
21. Genetta, T., Ruezinsky, D. & Kadesch, T. Displacement of an E-box-binding repressor by basic helix-loop-helix proteins: implications for B-cell specificity of the immunoglobulin heavy-chain enhancer. *Mol. Cell Biol.* **14**, 6153–6163 (1994).
22. Webber, A. L., Ingram, R. S., Levorse, J. M. & Tilghman, S. M. Location of enhancers is essential for the imprinting of H19 and Igf2 genes. *Nature* **391**, 711–715 (1998).
23. Leighton, P. A., Saam, J. R., Ingram, R. S., Stewart, C. L. & Tilghman, S. M. An enhancer deletion affects both H19 and Igf2 expression. *Genes Dev.* **9**, 2079–2089 (1995).
24. Dorsett, D. Distant liaisons: long-range enhancer-promoter interactions in *Drosophila*. *Curr. Opin. Genet. Dev.* **9**, 505–514 (1999).
25. Merli, C., Bergstrom, D. E., Cygan, J. A. & Blackman, R. K. Promoter specificity mediates the independent regulation of neighboring genes. *Genes Dev.* **10**, 1260–1270 (1996).
26. DiLeone, R. J., Russell, L. B. & Kingsley, D. M. An extensive 3' regulatory region controls expression of Bmp5 in specific anatomical structures of the mouse embryo. *Genetics* **148**, 401–408 (1998).
27. Small, S., Arnosti, D. N. & Levine, M. Spacing ensures autonomous expression of different stripe enhancers in the even-skipped promoter. *Development* **119**, 762–772 (1993).
28. Fujioaka, M., Emi-Sarker, Y., Yusibova, G. L., Goto, T. & Jaynes, J. B. Analysis of an even-skipped rescue transgene reveals both composite and discrete neuronal and early blastoderm enhancers, and multi-stripe positioning by gap gene repressor gradients. *Development* **126**, 2527–2538 (1999).
29. Mannervik, M., Nibu, Y., Zhang, H. & Levine, M. Transcriptional coregulators in development. *Science* **284**, 606–609 (1999).
30. Schroder, C., Tautz, D., Seifert, E. & Jackle, H. Differential regulation of the two transcripts from the *Drosophila* gap segmentation gene hunchback. *EMBO J.* **7**, 2881–2887 (1988).
31. Bell, A. C., West, A. G. & Felsenfeld, G. Insulators and boundaries: versatile regulatory elements in the eukaryotic genome. *Science* **291**, 447–450 (2001).
32. Kellum, R. & Schedl, P. A position-effect assay for boundaries of higher order chromosomal domains. *Cell* **64**, 941–950 (1991).
33. Choi, O. R. & Engel, J. D. Developmental regulation of beta-globin gene switching. *Cell* **55**, 17–26 (1988).
34. Ohtsuki, S., Levine, M. & Cai, H. N. Different core promoters possess distinct regulatory activities in the *Drosophila* embryo. *Genes Dev.* **12**, 547–556 (1998).
35. Butler, J. E. & Kadonaga, J. T. Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. *Genes Dev.* **15**, 2515–2519 (2001).
36. Perkins, A. C., Gaensler, K. M. & Orkin, S. H. Silencing of human fetal globin expression is impaired in the absence of the adult beta-globin gene activator protein EKLF. *Proc. Natl Acad. Sci. USA* **93**, 12267–12271 (1996).
37. Roeder, R. G. Role of general and gene-specific cofactors in the regulation of eukaryotic transcription. *Cold Spring Harb. Symp. Quant. Biol.* **63**, 201–218 (1998).
38. Freiman, R. N. *et al.* Requirement of tissue-selective TBP-associated factor TAFII105 in ovarian development. *Science* **293**, 2084–2087 (2001).
39. Hiller, M. A., Lin, T. Y., Wood, C. & Fuller, M. T. Developmental regulation of transcription by a tissue-specific TAF homolog. *Genes Dev.* **15**, 1021–1030 (2001).
40. Holmes, M. & Tjian, R. Promoter selective properties of the TBP-related factor TRF1. *Science* **288**, 867–870 (2000).
41. Rabenstein, M. D., Zhou, S., Lis, J. T. & Tjian, R. TATA box-binding protein (TBP)-related factor 2 (TRF2), a third member of the TBP family. *Proc. Natl Acad. Sci. USA* **96**, 4791–4796 (1999).
42. Kaltenbach, L., Horner, M. A., Rothman, J. H. & Mango, S. E. The TBP-like factor CcTLF is required to activate RNA polymerase II transcription during *C. elegans* embryogenesis. *Mol. Cell* **6**, 705–713 (2000).
43. Veenstra, G. J., Weeks, D. L. & Wolffe, A. P. Distinct roles for TBP and TBP-like factor in early embryonic gene transcription in *Xenopus*. *Science* **290**, 2312–2315 (2000).
44. Zhang, D., Penttila, T. L., Morris, P. L., Teichmann, M. & Roeder, R. G. Spermiogenesis deficiency in mice lacking the Trf2 gene. *Science* **292**, 1153–1155 (2001).
45. Martianov, I. *et al.* Late arrest of spermiogenesis and germ cell apoptosis in mice lacking the TBP-like TLF/TRF2 gene. *Mol. Cell* **7**, 509–515 (2001).
46. Hochheimer, A., Zhou, S., Zheng, S., Holmes, M. & Tjian, R. Promoter Selectivity and target gene identification of a DREF-containing TRF2 complex. *Nature* **420**, 439–445 (2002).
47. Kim, Y.-J., Bjorklund, S., Li, Y., Sayre, M. H. & Kornberg, R. D. A multiprotein mediator of transcriptional activation and its interaction with the C-terminal repeat domain of RNA polymerase II. *Cell* **77**, 599–608 (1994).
48. Koleske, A. J. & Young, R. A. An RNA polymerase II holoenzyme responsive to activators. *Nature* **368**, 466–469 (1994).
49. Fondell, J. D., Ge, H. & Roeder, R. G. Ligand induction of a transcriptionally active thyroid hormone receptor coactivator complex. *Proc. Natl Acad. Sci. USA* **93**, 8329–8333 (1996).
50. Rachez, C. *et al.* Ligand-dependent transcription activation by nuclear receptors requires the DRIP complex. *Nature* **398**, 824–828 (1999).
51. Gu, W. *et al.* A novel human SRB/MED-containing cofactor complex, SMCC, involved in transcription regulation. *Mol. Cell* **3**, 97–108 (1999).
52. Akoulitchev, S., Chuikov, S. & Reinberg, D. TFIID is negatively regulated by cdk8-containing mediator complexes. *Nature* **407**, 102–106 (2000).
53. Taatjes, D. J., Naar, A. M., Andel, F., Nogales, E. & Tjian, R. Structure, function, and activator-induced conformations of the CRSP co-activators. *Science* **295**, 1058–1062 (2002).
54. Naar, A. M. *et al.* Composite co-activator ARC mediates chromatin-directed transcriptional activation. *Nature* **398**, 828–832 (1999).
55. Ryu, S., Zhou, S., Ladurner, A. G. & Tjian, R. The transcriptional cofactor complex CRSP is required for activity of the enhancer-binding protein Sp1. *Nature* **397**, 446–450 (1999).
56. Ito, M. & Roeder, R. G. The TRAP/SMCC/Mediator complex and thyroid hormone receptor function. *Trends Endocrinol. Metab.* **12**, 127–134 (2001).
57. Dotson, M. R. *et al.* Structural organization of yeast and mammalian mediator complexes. *Proc. Natl Acad. Sci. USA* **97**, 14307–14310 (2000).
58. Malik, S. & Roeder, R. G. Transcriptional regulation through Mediator-like coactivators in yeast and metazoan cells. *Trends Biochem. Sci.* **25**, 277–283 (2000).
59. Glass, C. K. & Rosenfeld, M. G. The coregulator exchange in transcriptional functions of nuclear receptors. *Genes Dev.* **14**, 121–141 (2000).
60. Bromberg, J. & Chen, X. STAT proteins: signal transducers and activators of transcription. *Methods Enzymol.* **333**, 138–151 (2001).
61. Wang, W. *et al.* Diversity and specialization of mammalian SWI/SNF complexes. *Genes Dev.* **10**, 2117–2130 (1996).
62. Goodman, R. H. & Smolik, S. CBP/p300 in cell growth, transformation, and development. *Genes Dev.* **14**, 1553–1577 (2000).
63. Tamkun, J. W. *et al.* Brahma: a regulator of *Drosophila* homeotic genes structurally related to the yeast transcriptional activator SNF2/SWI2. *Cell* **68**, 561–572 (1992).
64. Khavari, P. A., Peterson, C. L., Tamkun, J. W., Mendel, D. B. & Crabtree, G. R. BRG1 contains a conserved domain of the SWI2/SNF2 family necessary for normal mitotic growth and transcription. *Nature* **366**, 170–174 (1993).
65. Feng, Q. & Zhang, Y. The NuRD complex: linking histone modification to nucleosome remodeling. *Curr. Top. Microbiol. Immunol.* **274**, 269–290 (2003).
66. Tsukiyama, T., Daniel, C., Tamkun, J. & Wu, C. ISWI, a member of the SWI2/SNF2 ATPase family, encodes the 140 kDa subunit of the nucleosome remodeling factor. *Cell* **83**, 1021–1026 (1995).
67. LeRoy, G., Loyola, A., Lane, W. S. & Reinberg, D. Purification and characterization of a human factor that assembles and remodels chromatin. *J. Biol. Chem.* **275**, 14787–14790 (2000).
68. Olave, I. A., Reck-Peterson, S. L. & Crabtree, G. R. Nuclear actin and actin-related proteins in chromatin remodeling. *Annu. Rev. Biochem.* **71**, 755–781 (2002).
69. Kehle, J. *et al.* dMi-2, a hunchback-interacting protein that functions in polycomb repression. *Science* **282**, 1897–1900 (1998).
70. Markstein, M., Markstein, P., Markstein, V. & Levine, M. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. *Proc. Natl Acad. Sci. USA* **99**, 763–768 (2002).
71. Berman, B. P. *et al.* Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl Acad. Sci. USA* **99**, 757–762 (2002).
72. Rebeiz, M., Reeves, N. L. & Posakony, J. W. SCORE: A computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. *Proc. Natl Acad. Sci. USA* (in the press).
73. Halfon, M. S., Grad, Y., Church, G. M. & Michelson, A. M. Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome Res.* **12**, 1019–1028 (2002).
74. Rajewsky, N., Vergassola, M., Gaul, U. & Siggia, E. D. Computational detection of genomic cis-regulatory modules applied to body patterning in the early *Drosophila* embryo. *BMC Bioinform.* **3**, 30 (2002).

Acknowledgements We thank Y. Nibu, A. Ladurner and J. Ziegelbauer for preparing the figures. We also thank D. Rio and L. Mirels for critically reviewing the manuscript. M.L. is supported by the NIH and R.T. is funded in part by a grant from the NIH.

Correspondence and requests for materials should be addressed to R.T. (jmlim@uclink4.berkeley.edu) or M.L. (mlevine@uclink4.berkeley.edu).

Deep convection in the Irminger Sea forced by the Greenland tip jet

Robert S. Pickart*, Michael A. Spall*, Mads Hvid Ribergaard†, G. W. K. Moore‡ & Ralph F. Milliff§

* Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA

† Danish Meteorological Institute, Copenhagen DK-2100, Denmark

‡ University of Toronto, Ontario M5s 1A1, Canada

§ Colorado Research Associates Division, NWSA, Boulder, Colorado 80301, USA

Open-ocean deep convection, one of the processes by which deep waters of the world's oceans are formed, is restricted to a small number of locations (for example, the Mediterranean and Labrador seas). Recently, the southwest Irminger Sea has been suggested as an additional location for open-ocean deep convection. The deep water formed in the Irminger Sea has the characteristic temperature and salinity of the water mass that fills the mid-depth North Atlantic Ocean, which had been believed to be formed entirely in the Labrador basin. Here we show that the most likely cause of the convection in the Irminger Sea is a low-level atmospheric jet known as the Greenland tip jet, which forms periodically in the lee of Cape Farewell, Greenland, and is associated with elevated heat flux and strong wind stress curl. Using a history of tip-jet events derived from meteorological land station data and a regional oceanic numerical model, we demonstrate that deep convection can occur in this region when the North Atlantic Oscillation Index is high, which is consistent with observations. This mechanism of convection in the Irminger Sea differs significantly from those known to operate in the Labrador and Mediterranean seas.

Nearly 100 years ago it was hypothesized that most of the deep water in the North Atlantic was formed in the southwest Irminger Sea by open-ocean convection¹. The surface waters of the Labrador Sea were deemed too fresh to permit wintertime overturning. Over the next few decades this idea was subject to much scrutiny, and a debate ensued in the scientific literature as to whether the Labrador Sea or the Irminger Sea was the main source of deep water to the North Atlantic. Despite evidence supporting both points of view^{2,3}, the notion that deep convection took place east of Greenland eventually fell out of favour. Today, it is commonly accepted that the body of weakly stratified water found at mid-depth (500–1,500 m) in the North Atlantic stems entirely from the Labrador Sea—hence the term ‘Labrador Sea Water’.

Recently, compelling evidence has reawoken interest in the original hypothesis⁴, suggesting that a significant amount of Labrador Sea Water may be formed outside the Labrador Sea after all. If true, this means that the meridional overturning circulation of the North Atlantic has a higher level of complexity than previously thought, which needs to be taken into account when interpreting observations and diagnosing models. For example, much has been made of the rapid appearance of newly ventilated Labrador Sea Water throughout the subpolar North Atlantic in the early 1990s (refs 5, 6); a second source of convection would cause us to revise the ideas put forward to explain this. One of the main puzzles regarding convection east of Greenland is the spatial scale of the overturning, which seems to be confined to the southern portion of the Irminger Sea⁴. Here we show that convection in the Irminger Sea is driven by the intermittent Greenland tip jet. This small-scale atmospheric jet results in large air–sea heat flux and strong wind stress curl, which together force deep oceanic mixing.

The Greenland tip jet

Part of the reason for the emphasis on the Labrador Sea as a convective site is the harsh wintertime atmospheric forcing there. Frigid polar air blows off the Canadian land mass as storms pass by, removing heat over a large portion of the Labrador Sea^{7,8}. This strong buoyancy loss is depicted clearly in the National Centers for Environmental Prediction (NCEP) reanalysis global meteorological fields, which have a spatial resolution of 1–2°. By contrast, the NCEP

fields show substantially lower heat loss in the Irminger Sea⁹. Recently, however, attention has been focused on a phenomenon known as the Greenland tip jet, a narrow, low-level jet that periodically develops in the lee of Cape Farewell^{10,11}. It forms when high-level northwesterly winds descend orographically on the eastern (leeward) side of Greenland and accelerate owing to conservation of Bernoulli function, drawing cold air over the southern Irminger Sea¹⁰. Until now, the frequency of such tip-jet events—and their effect on the ocean—has not been investigated.

In the winter of 1996–7, a set of field programmes was undertaken to study wintertime processes in the western North Atlantic^{12,13}. This included the implementation of the Coupled Ocean–Atmosphere Mesoscale Prediction System (COAMPS) model¹⁴. The model was run at 45-km resolution over a domain that included the Labrador and Irminger seas, and the heat flux fields were subsequently corrected for inaccuracies in high-wind conditions using *in situ* data from the field measurements¹⁵. The COAMPS model successfully captured the tip-jet occurrences during that winter, and, in particular, one robust event from 18 to 22 February. We begin our analysis by focusing on this event (which was also the subject of a previous modelling study¹⁰), in order to familiarize the reader with the phenomenon, and to investigate the effect of the tip jet on the underlying ocean.

The COAMPS zonal 10-m wind speed and 2-m air temperature (Fig. 1a) reveal the scales of the tip jet. It spans less than 2° of latitude (which is roughly the meridional grid spacing of the NCEP global data assimilation system data set), drawing cold air over the southern Irminger Sea in a fairly narrow tongue. This small meridional scale results in very strong localized wind stress curl (Fig. 1b). Note that the asymmetry of the jet leads to stronger cyclonic curl on the northern side of the jet axis. How does the COAMPS representation of this phenomenon compare to observations? To answer this we analysed wind data from the National Aeronautics and Space Administration scatterometer (NSCAT)¹⁶. The February 1997 tip-jet event is even stronger in the satellite data, where surface wind vectors are reported at 25-km resolution. The meridional scale is the same as in the model, and the cyclonic wind stress curl is of the order of 10^{-4} N m^{-3} (Fig. 1c). This is nearly three orders of magnitude stronger than the large-scale wintertime wind

stress curl over the western subpolar North Atlantic. It should be noted that the same tip-jet event is nearly indiscernible in the NCEP data, appearing only briefly on 18 February. By contrast, the tip-jet signature dominates the month-long average in the high-resolution COAMPS and NSCAT fields (there are other less-intense tip-jet events that month).

The strong winds and cold temperatures of the Greenland tip jet lead to enhanced heat loss in the southern Irminger Sea along a band centred near 60° N. According to the COAMPS model, the amplitude of the flux of sensible heat + latent heat averaged over the entire event is 475 W m^{-2} . This is comparable to the observed heat

flux during storm events in the western Labrador Sea during the same winter (when deep convection was observed¹⁵). The NCEP heat flux for the February 1997 event is only 225 W m^{-2} , which is indistinguishable from background values. Hence, the global weather models do not capture this phenomenon. The next question is, does the Greenland tip jet develop often enough to affect the Irminger Sea over an entire winter?

High-frequency forcing

A system of meteorological weather stations has been maintained along the coast of Greenland for several decades¹¹, and provides the opportunity to construct a climatology of tip-jet events. We must first show, however, that these events are captured in the weather station data. The most appropriately positioned station is Prins Christian Sund (labelled P in Fig. 1), which clearly captured the February 1997 event (Fig. 2). By contrast, stations I and N, just north of Cape Farewell, showed no signs of the jet even though the same low-pressure system was measured at these locations (Fig. 2a). Station A (western Cape Farewell) showed a moderate jet signal, and station Q (northwest Cape) showed an even weaker signature. Hence the cluster of meteorological stations in southern Greenland agree with the COAMPS depiction of the February 1997 tip-jet

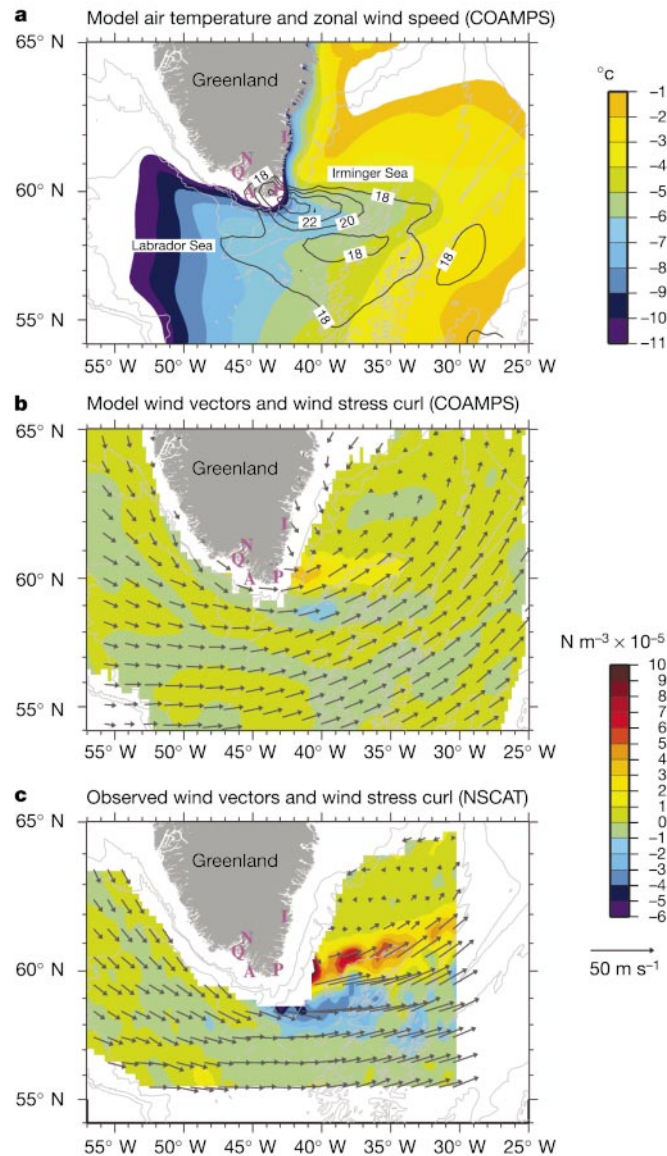


Figure 1 The Greenland tip jet depicted by the COAMPS weather model and observed by satellite. **a**, Average 10-m zonal wind speed $\geq 18 \text{ m s}^{-1}$ (contours), and 2-m air temperature (colour) on 18 February 1997, from COAMPS. The bottom depth contours (grey) are 1,000 m, 2,000 m and 3,000 m. The locations of the Greenland meteorological stations used in the study are denoted by the magenta lettering (P, Prins Christian Sund; A, Angisoq; Q, Qaqortoq; N, Narsarsuaq; I, Ikermiarsuk). **b**, COAMPS 10-m wind vectors and wind stress curl (colour) for 18 February 1997. Only every other wind vector is plotted. **c**, Average surface wind vectors and wind stress curl (colour) measured by NSCAT over the period 18–22 February 1997. The raw data were averaged into 0.5° bins, following previous methodology²⁴ (only bins with ≥ 2 realizations were used). Every other vector is plotted.

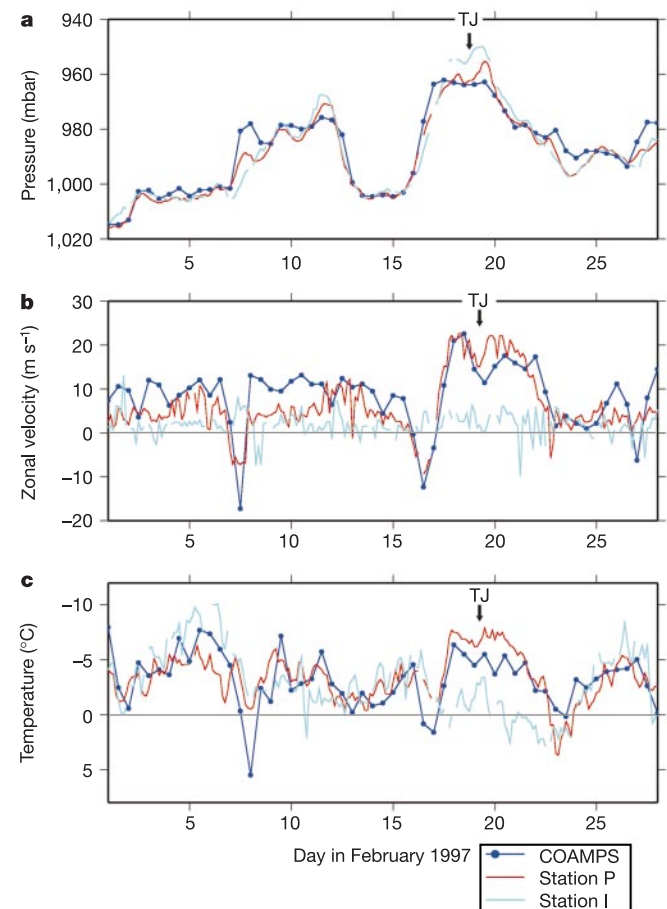


Figure 2 A tip-jet event recorded by Greenlandic meteorological land stations. The COAMPS model output (averaged over the region 59–60°N, 37–42°W) is compared with the observed meteorological time series (see key) for February 1997. The land station data are recorded every 3 hours (gaps indicate missing data); see Fig. 1 for the locations of the meteorological stations. The tip-jet event is denoted by TJ. **a**, Sea-level pressure. **b**, Zonal 10-m wind (positive is westerly). The COAMPS and station P winds are significantly correlated ($r = 0.72$); neither of them is correlated with the station I wind record. **c**, 2-m air temperature. The two meteorological station temperature records have been offset by 4° (warmer) for plotting purposes.

event, lending credence to the small meridional scale of the feature.

Inspection of the meteorological time series at station P indicates that tip-jet events are easily recognizable as bursts of strong westerly winds, accompanied by anomalously low air temperatures and sea-level pressure (for example, Fig. 2). Accordingly, we used empirical orthogonal function (EOF) analysis to identify objectively the number of tip-jet events each winter from 1962 to 1997. Maximum correlation resulted when the pressure record was lagged by 0.5 days

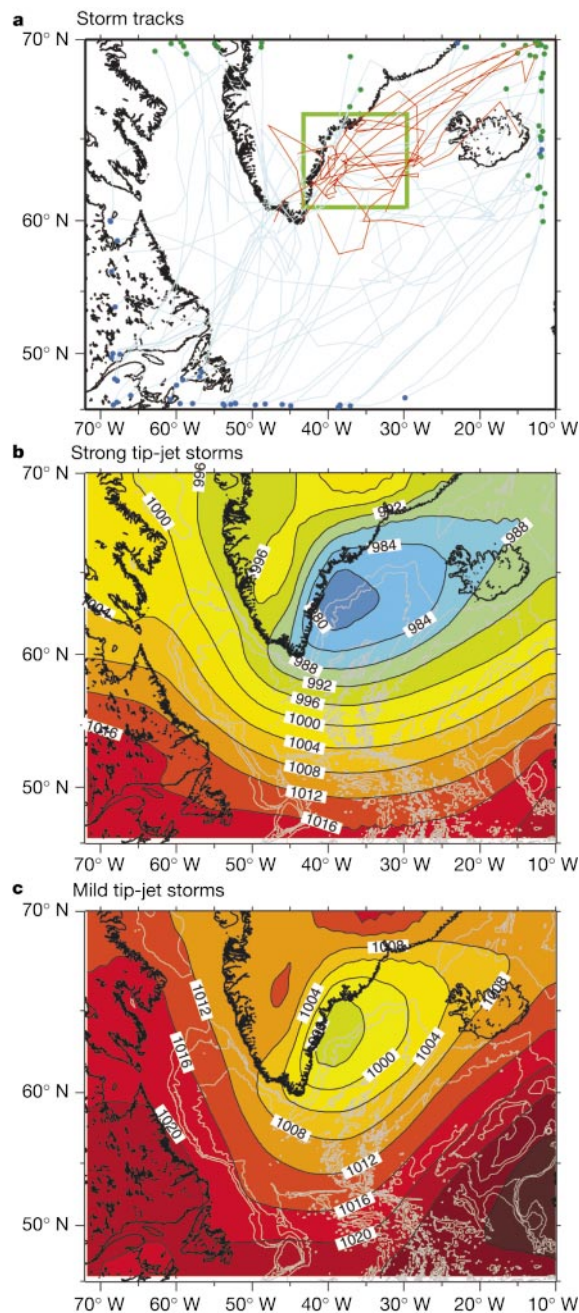


Figure 3 Triggering of tip-jet events by passing storms. **a**, Tracks of all the low-pressure systems during winter 1993 (December–March), computed from the 6-hourly NCEP fields. A blue (green) circle denotes where the storm entered (left) the domain. The majority of the storms traversed the domain from southwest to northeast. The red-coloured portions of the tracks denote those times when the tip jet was observed at station P. The green box is the ‘trigger box’ (see text). **b**, Average NCEP sea-level pressure (mbar) for the periods when strong tip-jet storms were located in the trigger box. The bottom depth contours (grey) are 1,000 m, 2,000 m and 3,000 m. **c**, Same as **b** for the mild tip-jet storms.

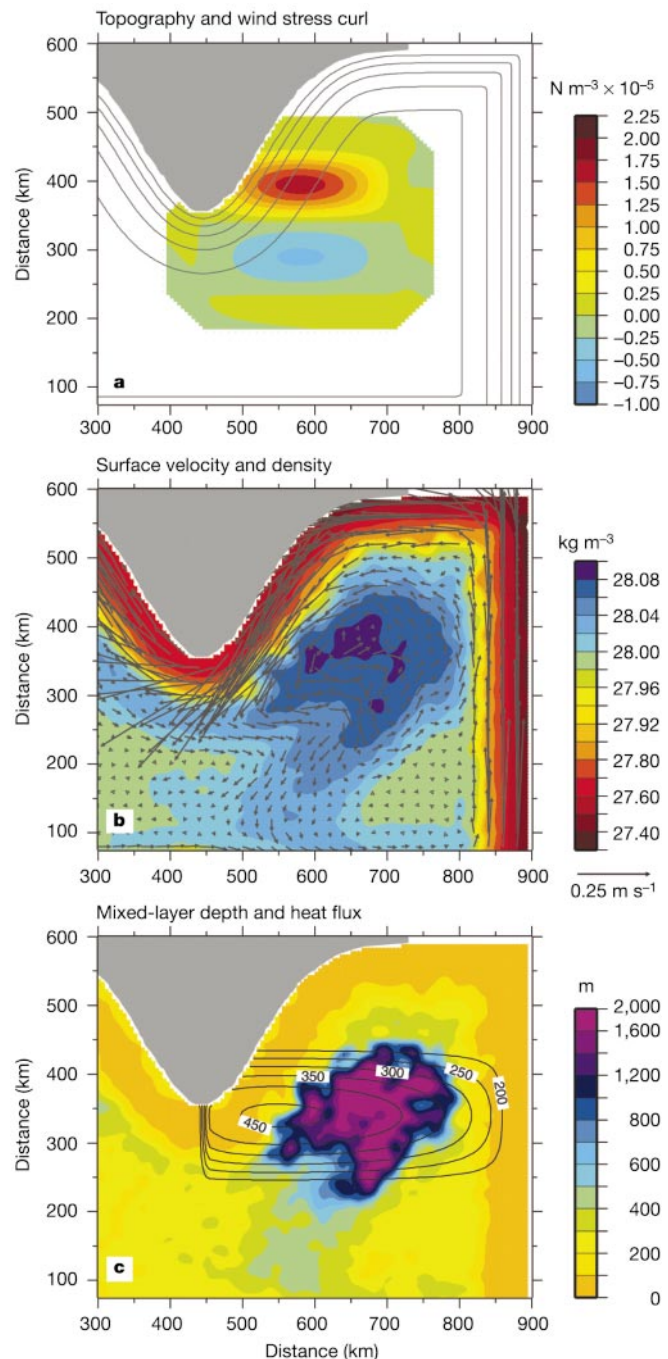


Figure 4 Tip-jet-forced oceanic convection in a regional numerical model. **a**, Configuration and idealized wind stress curl forcing of the oceanic primitive equation model. The full domain extends from 0 to 900 km in the zonal direction, and from 0 to 600 km in the meridional direction. The horizontal resolution is 5 km, and there are 20 levels in the vertical. The bathymetric contours are 1,000–3,000 m in 500-m increments. The magnitude of the curl corresponds to the average COAMPS value over the 18–22 February 1997 tip-jet event. Although this is significantly smaller than observed by NSCAT (Fig. 1), the event was larger than normal; hence the COAMPS value is deemed more representative of typical tip-jet conditions. **b**, Surface velocity vectors (every fourth point) overlaid on surface density at the end of the second winter. The vectors represent the average over the last 30 days of the integration. Both the density and velocity were smoothed using a laplacian spline filter to remove small-scale noise. **c**, Final mixed-layer depth (smoothed as in **b**) in relation to the idealized tip-jet heat flux forcing (contours in W m^{-2}).

(Fig. 2), and the time series were first low-passed using a 20-hour filter. The dominant EOF of the combined pressure, zonal velocity, and air temperature time series on average accounted for 55% of the total variance, more than twice that of the other two modes. This mode clearly represents the variability associated with tip-jet events. We defined a major event as one in which the reconstructed zonal EOF velocity exceeded 9 m s^{-1} (according to this criterion, the 18–22 February 1997 event was the only major one that month). Results were not sensitive to this choice—the tip-jet events stand out clearly in the original data as well as the EOFs.

We tabulated occurrences of the tip jet over the winter months (December–March) at station P. Although February tends to have more events, there is little variation throughout the winter. On average, a tip-jet event occurs roughly every 10 days, each event lasting approximately three days. The average zonal wind speed is 12.4 m s^{-1} (note that station P is upstream of the maximum jet amplitude, Fig. 1a). Thus, this phenomenon is extremely common. It is in fact one of the dominant winter weather patterns over southern Greenland (L. Rasmussen, personal communication). In addition, there is a ‘shoulder season’ in late autumn and early spring, during which additional tip-jet events often occur.

Is there a clear relationship between storms in the vicinity of Greenland and the occurrence of the tip jet? To investigate this, we tracked all the low-pressure systems during the winter of 1993 (a strong winter) using the NCEP data (which captures synoptic systems well), and compared them to the station P time series. This reveals that the tip jet tends to form when storms are situated between Cape Farewell and Iceland (Fig. 3a). In particular, all 10 events in 1993 occurred when the centre of the low-pressure system passed through the ‘trigger box’ delineated in Fig. 3a. There were seven other storms that passed through this box, and inspection of the EOF time series from station P reveals that every one of them resulted in a mild tip-jet event (that is, not strong enough to meet the criterion specified above). This is because the storms in question were not as robust (compare Fig. 3b and c). Hence there can be many tip-jet events (both strong and mild) during a robust winter; these events—apparently without exception—are triggered when a storm passes to the northeast of Cape Farewell.

Although on average there is little intra-seasonal variation in tip-jet occurrences, there is strong inter-annual variability: some years have only a few events (in fact two of the winters during the

study period had no major events). Storm activity in the subpolar North Atlantic is strongly related to the North Atlantic Oscillation (NAO) wintertime index¹⁷. A positive index is associated with a more northeasterly storm track, and so a greater number of synoptic systems pass near Greenland¹⁸. This implies that there should be a relationship between the NAO and the number of tip-jet events in a winter. Our data reveal a statistically significant correlation over the 35-year time period, with a larger index implying more frequent events. Hence, a positive NAO tends to favour the occurrence of convective overturning in the southwest Irminger Sea, as it does in the Labrador Sea¹⁹. However, tip-jet events seem to be also affected by more subtle changes in storm trajectories. In winter of 1999 the centre of action of the NAO shifted to the east²⁰, and the trajectories of numerous storms passed farther offshore of Greenland. Although the NAO index was high for that winter, the eastward shift resulted in more moderate conditions in the Labrador Sea (and less tendency for deep convection there²⁰). In 1999 there were 10 major tip-jet events, typical for a high-NAO winter. However, the eastward shift in storm trajectories resulted in significantly warmer air being advected over the Irminger Sea within the tip jet (on average 4°C warmer than for the winter of 1993). This corresponds to a reduction in the sensible heat loss of roughly 115 W m^{-2} . Hence, as for the Labrador Sea, such an anomalous NAO state is apparently less favourable for deep overturning east of Greenland as well.

Oceanic response

To investigate the effect of tip-jet forcing on the ocean, and in particular its potential to drive deep convection in the Irminger Sea, we ran a regional application of the MIT primitive equation numerical model²¹. The domain, meant to correspond to the western subpolar North Atlantic, has an idealized representation of Greenland and the continental slope (Fig. 4a). The model was initialized with measured temperature and salinity profiles from the Irminger Sea, representative of the stratification in late summer found during a high-NAO year. Density is calculated from temperature and salinity using a linear equation of state. In order to include the effects of the buoyant boundary current encircling the Irminger Sea (the Irminger Current), the stratification and horizontal velocity are restored towards a narrow, surface-intensified geostrophic jet along the southern and eastern boundaries. The model boundary current follows the topographic contours cyclonically around the basin, passing along the east coast of Greenland.

We forced the model with a representation of the Greenland tip jet. The climatology indicates that for a typical high-NAO winter there are 10 major events from mid-December to mid-March. As there are often significant events in late autumn and early spring, as well as numerous mild events throughout the winter, we parameterized this by applying 14 equally strong events over 120 days (each event lasts 4 days and they are spaced 8 days apart). Results are not overly sensitive to these details. The model was initially run for one year using the wind forcing only, allowing the oceanic circulation to spin up. This provides an important pre-conditioning effect for the subsequent winter in which both wind and buoyancy forcing are applied.

We present results from the end of the second winter, during which both the tip-jet wind stress curl (Fig. 4a) and heat flux (Fig. 4c) were applied. A strong cyclonic gyre develops to the east of Cape Farewell, with a weaker anti-cyclonic swirl to the south (Fig. 4b); this is to be expected from the wind forcing. A region of increased surface density and deep mixed-layers, roughly 200 km in diameter, forms as well to the east of Cape Farewell (Fig. 4b, c). The deepest mixed layers, which took roughly 2.5 months to develop, extend to 2,000 m. This is clearly deep convection, in line with recent hydrographic observations⁴ that indicate overturning to 1,800 m in this area. The location and size of the model convective patch agree closely with that suggested by the measured potential vorticity (PV) distribution during the high-NAO period of the

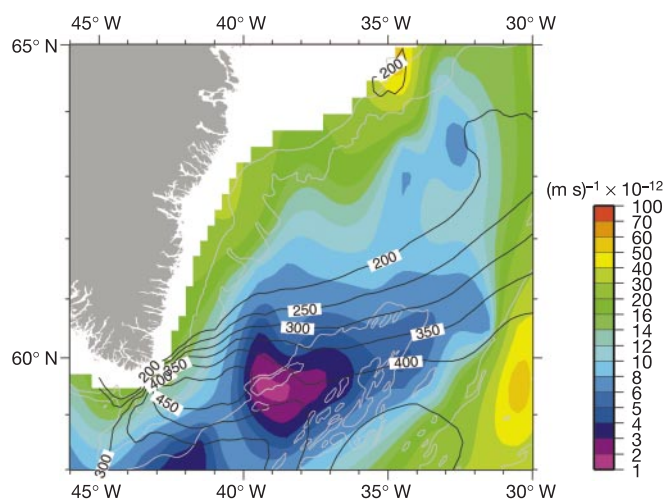


Figure 5 Observational support for tip-jet-forced deep convection east of Greenland. The figure shows average planetary potential vorticity (colour) at 1,000 m for the period 1989–97, from a climatological database⁴, in relation to the COAMPS tip-jet heat flux forcing (18–22 February 1997). This represents the observational analogue of Fig. 4c. The bottom depth contours (grey) are 1,000 m, 2,000 m and 3,000 m.

1990s⁴ (compare Figs 4c and 5). Both fields indicate a well-defined, isolated area of overturning in the southern Irminger Sea.

The wind forcing provided by the tip jet significantly affects the overturning. Strong Ekman suction decreases the thickness of the warm surface layer in the region of buoyancy loss, reducing the amount of heat required to be removed during winter. Calculations with cooling but no wind stress show that the deep convection occurs significantly later in the season, and the final volume of ventilated water is reduced by approximately 35%. An additional calculation was done in which a shallow layer of fresh water was included near the coast, representative of the East Greenland Current (initialized with a measured near-shore hydrographic profile). It produced convection to 2,000 m with a total volume nearly as large as the standard case. This insensitivity is because most of the fresh water remains trapped to the boundary, as is observed in this region⁴. Finally, convection reaches only 700 m when the initial density profile is representative of the stratification during low-NAO conditions. This suggests that deep overturning is unlikely to occur in low-NAO winters, or even during the first winter of strong forcing following several low-NAO years. This is consistent with the notion that the Irminger Sea has a convective memory longer than one year⁴.

Importance of tip-jet-forced convection

The circumstances surrounding convection in the Irminger Sea are very different from those in the other two locations of deep overturning in the North Atlantic—the Labrador and Mediterranean seas. The Labrador Sea has broad-scale wind and buoyancy forcing over most of the basin, with no strong wind stress curl. The small spatial scale of the forcing in the Irminger Sea is more reminiscent of the Mediterranean Sea, yet there are fundamental differences. The mistral and tramontane wind events in the Mediterranean occur less often, and can last substantially longer²². Furthermore, these winds are meridional (northerly). A modelling study of the Mediterranean case—using sustained winds over several months—demonstrated that a single cyclonic gyre develops directly underneath the forcing region, within which the deepest convection occurs²³. By contrast, in the Irminger Sea the zonal orientation of the (highly intermittent) tip jet spins up a dipole circulation, which resides to the east of the strongest forcing. Diagnosis of a series of model calculations indicates that this offset is due to the advection of density by the wind-driven flow. The effect of the anti-cyclonic circulation is evident by the tongue of dense water extending to the southwest (Fig. 4b), which enables deep convection to occur over a broader region (Fig. 4c). Both the model and observations show the eastward displacement of the Irminger Sea convective patch relative to the forcing.

It is important to note that we are not advocating the Irminger Sea as the sole or primary source of Labrador Sea Water. It does, however, appear to be a significant source. Although there are no direct estimates of new water production in the Irminger Sea, the volume of new water formed in the model (corresponding to the region where mixed-layer depths exceed 1,000 m) is $4.5 \times 10^{13} \text{ m}^3$. Using the $PV = 5 \text{ m}^{-1} \text{ s}^{-1} \times 10^{-12}$ contour as the delimiter of newly convected water in the observations (Fig. 5), and assuming an average convective depth of 1,500 m, gives a similar estimate of $4.6 \times 10^{13} \text{ m}^3$. This is comparable to the measured amount of new water formed in the Labrador Sea in the winter of 1996–7 ($4.0 \times 10^{13} \text{ m}^3$, the only such observational estimate for the Labrador Sea). Admittedly that winter was moderate⁹; none the less, it suggests that the Irminger Sea source is important. Furthermore, because the convected water formed in the Irminger Sea is not

confined within the cyclonic gyre (Fig. 4), it can more readily affect the surrounding area.

The implications of an additional source of Labrador Sea Water are far-reaching. It alters our view of the mid-depth component of the North Atlantic meridional overturning circulation, and means that we need to consider both observations and modelling studies from a new perspective. Our results also provide a striking example of the coupling that can occur between high-frequency, small-scale events in the atmosphere, and low-frequency, climatically important processes in the ocean. □

Received 25 October 2002; accepted 19 March 2003; doi:10.1038/nature01729.

1. Nansen, F. Das Bodenwasser und die Abkühlung des Meeres. *Int. Rev. Ges. Hydrobiol. Hydrogr.* **V**, 1–42 (1912).
2. Wüst, G. Der subarktische Bodenstrom in der westatlantischen Mulde. *Ann. Hydrogr. Marit. Meteorol.* **IV/VI**, 249–256 (1943).
3. Nielsen, J. *Greenland. The Discovery of Greenland, Exploration, and the Nature of the Country* Vol. I, 185–230 (C.A. Reitzel, Copenhagen, 1928).
4. Pickart, R. S., Straneo, F. & Moore, G. W. K. Is Labrador Sea Water formed in the Irminger Basin? *Deep Sea Res.* **150**, 23–52 (2003).
5. Sy, A. et al. Surprisingly rapid spreading of newly formed intermediate waters across the North Atlantic Ocean. *Nature* **386**, 675–679 (1997).
6. Rhein, M. et al. Labrador Sea water: Pathways, CFC-inventory, and formation rates. *J. Phys. Oceanogr.* **32**, 648–665 (2002).
7. Bumpke, K., Karger, U. & Uhlig, K. Measurements of turbulent fluxes of momentum and sensible heat over the Labrador Sea. *J. Phys. Oceanogr.* **32**, 401–410 (2002).
8. Moore, G. W. K. & Renfrew, I. A. An assessment of the surface turbulent heat fluxes from the NCEP-NCAR reanalysis over the western boundary currents. *J. Clim.* **15**, 2020–2037 (2002).
9. Pickart, R. S., Torres, D. J. & Clarke, R. A. Hydrography of the Labrador Sea during active convection. *J. Phys. Oceanogr.* **32**, 428–457 (2002).
10. Doyle, J. D. & Shapiro, M. A. Flow response to large-scale topography: the Greenland tip jet. *Tellus* **51**, 728–748 (1999).
11. Cappelen, J., Jørgensen, B. V., Laursen, E. V., Stannius, L. S. & Thomsen, R. S. *The Observed Climate of Greenland, 1958–99, with Climatological Standard Normals, 1961–90* (Technical Report 00-18, Danish Meteorological Institute, Copenhagen, 2001).
12. Labsea Group. The Labrador Sea deep convection experiment. *Bull. Am. Meteorol. Soc.* **79**, 2033–2058 (1998).
13. Browning, K. A., Chalon, J.-P. & Thorpe, A. J. The Fronts and Atlantic Storm-Track Experiment (FASTEX). *Q. J. R. Meteorol. Soc.* **125**, 3129–3617 (1999).
14. Renfrew, I. A., Moore, G. W. K., Holt, T. R., Chang, S. W. & Guest, P. Mesoscale forecasting during a field program: Meteorological support of the Labrador Sea deep convection experiment. *Bull. Am. Meteorol. Soc.* **80**, 605–620 (1999).
15. Renfrew, I. A., Moore, G. W. K., Guest, P. S. & Bumpke, K. A comparison of surface-layer heat flux and surface momentum flux observations over the Labrador Sea with ECMWF analyses and NCEP reanalyses. *J. Phys. Oceanogr.* **32**, 383–400 (2002).
16. Naderi, F. M., Freilich, M. H. & Long, D. G. Spaceborne radar measurement of wind velocity over the ocean—An overview of the NSCAT scatterometer system. *IEEE Proc.* **79**, 850–866 (1991).
17. Hurrell, J. W. Decadal trends in the North Atlantic Oscillation regional temperatures and precipitation. *Science* **269**, 676–679 (1995).
18. Rogers, J. C. Patterns of low-frequency monthly sea level pressure variability (1899–1986) and associated wave cyclone frequencies. *J. Clim.* **3**, 1364–1379 (1990).
19. Dickson, R. R., Lazier, J. R. N., Meincke, J., Rhines, P. B. & Swift, J. Long-term coordinated changes in the convective activity of the North Atlantic. *Prog. Oceanogr.* **38**, 241–295 (1996).
20. Dickson, R. R. & Meincke, J. in *Trans. ICES Symp. Hydrobiological Variability in the ICES Area, 1990–99* (ed. Turrell, William) (ICES Marine Science Symposium Series, Copenhagen, 2003).
21. Marshall, J., Hill, C., Perelman, L. & Adcroft, A. Hydrostatic, quasi-hydrostatic, and non-hydrostatic ocean modeling. *J. Geophys. Res.* **102**, 5733–5752 (1997).
22. Schott, F. et al. Observations of deep convection in the Gulf of Lions, Northern Mediterranean, during the winter of 1991/92. *J. Phys. Oceanogr.* **26**, 505–524 (1996).
23. Madec, G., Lott, F., Delecluse, P. & Crepon, M. Large-scale preconditioning of deep-water formation in the Northwestern Mediterranean Sea. *J. Phys. Oceanogr.* **26**, 1393–1408 (1996).
24. Milliff, R. F. & Morzel, J. The global distribution of the time-averaged wind stress curl from NSCAT. *J. Atmos. Sci.* **58**, 109–131 (2001).

Acknowledgements We thank L. Rasmussen for discussions about the meteorology around Greenland. The NSCAT surface winds were derived from the KU2000 retrieval provided by Remote Sensing Systems Inc. Support for this work was provided by the Ocean Sciences Division of the National Science Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.P. (rpickart@whoi.edu).

The DNA sequence of human chromosome 7

LaDeana W. Hillier*, Robert S. Fulton*, Lucinda A. Fulton*, Tina A. Graves*, Kymberlie H. Pepin*, Caryn Wagner-McPherson*, Dan Layman*, Jason Maas*, Sara Jaeger*, Rebecca Walker*, Kristine Wylie*, Mandeep Sekhon*, Michael C. Becker*, Michelle D. O'Laughlin*, Mark E. Schaller*, Ginger A. Fewell*, Kimberly D. Delehaunty*, Tracie L. Miner*, William E. Nash*, Matt Cordes*, Hui Du*, Hui Sun*, Jennifer Edwards*, Holland Bradshaw-Cordum*, Johar Ali*, Stephanie Andrews*, Amber Isak*, Andrew VanBrunt*, Christine Nguyen*, Feiyu Du*, Betty Lamar*, Laura Courtney*, Joelle Kalicki*, Philip Ozersky*, Lauren Bielicki*, Kelsi Scott*, Andrea Holmes*, Richard Harkins*, Anthony Harris*, Cynthia Madsen Strong*, Shunfang Hou*, Chad Tomlinson*, Sara Dauphin-Kohlberg*, Amy Kozlowicz-Reilly*, Shawn Leonard*, Theresa Rohlfing*, Susan M. Rock*, Aye-Mon Tin-Wollam*, Amanda Abbott*, Patrick Minx*, Rachel Maupin*, Catrina Strowmatt*, Phil Latreille*, Nancy Miller*, Doug Johnson*, Jennifer Murray*, Jeffrey P. Woessner*, Michael C. Wendl*, Shiaw-Pyng Yang*, Brian R. Schultz*, John W. Wallis*, John Spieth*, Tamberlyn A. Bieri*, Joanne O. Nelson*, Nicolas Berkowicz*, Patricia E. Wohldmann*, Lisa L. Cook*, Matthew T. Hickenbotham*, James Eldred*, Donald Williams*, Joseph A. Bedell*, Elaine R. Mardis*, Sandra W. Clifton*, Stephanie L. Chisoe*, Marco A. Marra*††, Christopher Raymond†, Eric Haugen†, Will Gillett†, Yang Zhou†, Rose James†, Karen Phelps†, Shawn Iadonato†, Kerry Bubb†, Elizabeth Simms†, Ruth Levy†, James Clendenning†, Rajinder Kaul†, W. James Kent‡, Terrence S. Furey‡, Robert A. Baertsch‡, Michael R. Brent§, Evan Keibler§, Paul Flicek§, Peer Bork||, Mikita Suyama||, Jeffrey A. Bailey¶, Matthew E. Portnoy#, David Torrents||, Asif T. Chinwalla*, Warren R. Gish*, Sean R. Eddy☆, John D. McPherson*††, Maynard V. Olson†, Evan E. Eichler¶, Eric D. Green#, Robert H. Waterston*†† & Richard K. Wilson*

* Genome Sequencing Center, Washington University School of Medicine, Campus Box 8501, 4444 Forest Park Avenue, St Louis, Missouri 63108, USA

† University of Washington Genome Center, 2225 Fluke Hall on Mason Road, Campus Box 352 145 Seattle, Washington 98195, USA

‡ Center for Biomolecular Science and Engineering, University of California, 321 BE Santa Cruz, California 95064, USA

§ Department of Computer Science, Washington University, Box 1045, St Louis, Missouri 63130, USA

|| EMBL, Meyerhofstrasse 1, Heidelberg 69117, Germany

¶ Department of Genetics, Center for Computational Genomics and Center for Human Genetics, Case Western Reserve University School of Medicine and University Hospitals of Cleveland, Cleveland, Ohio 44106, USA

Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Building 50, Room 5222, South Drive, Bethesda, Maryland 20892, USA

☆ Howard Hughes Medical Institute and Department of Genetics, Washington University School of Medicine, Campus Box 8232, 4566 Scott Ave. St Louis, Missouri 63110, USA

Human chromosome 7 has historically received prominent attention in the human genetics community, primarily related to the search for the cystic fibrosis gene and the frequent cytogenetic changes associated with various forms of cancer. Here we present more than 153 million base pairs representing 99.4% of the euchromatic sequence of chromosome 7, the first metacentric chromosome completed so far. The sequence has excellent concordance with previously established physical and genetic maps, and it exhibits an unusual amount of segmentally duplicated sequence (8.2%), with marked differences between the two arms. Our initial analyses have identified 1,150 protein-coding genes, 605 of which have been confirmed by complementary DNA sequences, and an additional 941 pseudogenes. Of genes confirmed by transcript sequences, some are polymorphic for mutations that disrupt the reading frame.

As the reference human genome sequence nears completion, the sequences of individual chromosomes are providing foundational information for genome structure, organization and evolution. Previously, chromosome 7 has been the focal point in the search for the gene associated with cystic fibrosis¹ and the frequent cytogenetic changes associated with some forms of cancer². Here, we describe our analysis of the chromosome 7 sequence, which has exploited orthology with the mouse genome in refining gene predictions and has discovered some unusual structural features that have been implicated in genetic diseases. The unrestricted access of the community to this sequence as it was generated has aided in the discovery of genes on chromosome 7 related to human health and well being, and we anticipate that the full sequence will provide further impetus to these studies.

General features of the chromosome 7 sequence

We generated the sequence of human chromosome 7 using a clone-by-clone shotgun sequencing strategy^{3,4} and organized the resulting sequence into 11 contigs (Fig. 1). With the exception of the centromere and one gap near the terminal end of the long arm (pter), the distances between contigs are relatively small (Table 1), with most sized by fluorescence *in situ* hybridization (FISH) of DNA fibres or by comparison to the analogous region in the mouse genome sequence. The DNA in these remaining gaps is repetitive or has proved recalcitrant to isolation in bacterial- or yeast-based cloning systems, including the screening of a series of large-insert genomic libraries that together provided 100-fold coverage of the human genome. On the basis of the size estimates of these gaps, the available sequence represents greater than 99.4% of the total euchromatic sequence.

Isolation of clones containing the telomeric and pericentromeric regions of the p arm were particularly hampered by the presence of repetitive sequences. A 7p-telomere-containing 'half-YAC'⁵ (yeast artificial chromosome) has been identified, but has been too unstable to sequence. However, three cosmid clones estimated to

†† Present addresses: Genome Sciences Centre, British Columbia Cancer Agency, 600 West 10th Avenue, Room 3427, Vancouver, British Columbia, V5Z-4E6, Canada (M.A.M.); Department of Genome Sciences, Box 357730, University of Washington, 1705 NE Pacific Street, Seattle, Washington 98195-7730, USA (R.H.W.); Baylor College of Medicine, 1 Baylor Plaza, Human Genome Sequencing Center, N1519, Houston, Texas 77030, USA (J.D.M.).

Table 1 Contiguous sequence lengths and gap sizes

Map	Sequence length (Mb)*	Gap size by FISH (kb)	Gap size by mouse (kb)†	Accession (NT_)	Comment‡
Gap	—	—	—	—	~34-kb telomeric gap
1	0.3	—	—	029998.5	—
Gap	—	NA	334	—	—
2	47.4	—	—	007819.11	Extended by 70 kb
Gap	—	50	54	—	—
3	2.1	—	—	030008.5	—
Gap	—	40	45	—	—
4	6.5	—	—	033968.2	Merged, ends in α-satellite
				034884.1	—
Gap	—	20–50	NA	—	—
5	0.8	—	—	023629.10	Centromeric, extended by 300 kb
Gap	—	NA	NA	—	—
6	0.2	—	—	023603.4	Unanchored centromeric (uc) clone, 190 kb uc clone added
Gap	—	NA	NA	—	—
7	12.7	—	—	007758.8	Right end extends into WBS region§
				034886.2	—
Gap	—	NA	NA	—	—
8	64.4	—	—	007933.10	Left end in WBS
Gap	—	NA	87	—	—
9	14.8	—	—	007914.10	—
Gap	—	100	99	—	—
10	0.7	—	—	034885.1	—
Gap	—	80	80	—	—
11	3.9	—	—	007741.10	29-kb gap closed
				028233.7	—

Values were created by FISH and estimation from orthologous mouse sequence. NA, not applicable.
*Status as of 1 July 2002, as reflected in NCBI build 31. These data were used for analyses presented here.
†Estimates of gap sizes based on mouse were obtained from A. Cook, M. Kamal and M. Zody (personal communication).
‡Includes any changes to the sequence since 1 July 2002 from continuing efforts to close all remaining gaps.
§Mouse genome lacks analogous region⁹.

be 34 kilobases (kb) from the true telomere were recovered and contain scattered telomere-specific repeats. Such difficulties were not encountered with the 7q telomeres, where the sequence appears to extend into the telomere. On either side of the centromere, we analysed the sequence for higher-order alpha satellite repeats indicative of centromeric boundaries. On the short (q) arm side of the centromere, the sequence clearly contains such boundary features. The boundary was less clear for the p-arm side of the centromere, and the search for it was compounded by the duplication of this region elsewhere on the chromosome.

The quality of the chromosome 7 sequence exceeds the 99.99% accuracy standard established by the International Human Genome Consortium for sequencing the human genome⁶. We further checked the integrity of the sequence and its assembly in two ways. First, as each clone was finished, an *in silico* digest of the sequence was compared to restriction digests of the clone DNA. We also checked the fully assembled sequence by performing *in silico* digests of clone-sized fragments across the chromosome against the underlying fingerprint data used to construct the physical map. In this way, we confirmed more than 99.9% of the testable bands.

Comparison to physical and genetic maps

We evaluated the completeness of the chromosome 7 sequence by looking for its representation of sequence-tagged sites (STSs) from previously constructed physical and genetic maps of the chromosome, specifically a YAC-based STS-content map⁷, the Genethon microsatellite-based genetic map⁸, and a chromosome-7-specific radiation-hybrid (RH) map⁷. There were only a small number of unassigned STSs (see Supplementary Information), which included those from multi-copy sequences, sequence polymorphisms, regions within the remaining chromosome 7 sequence gaps, and clerical errors that preclude accurate matching of STS names to their true underlying sequence.

We also used these maps to evaluate the assembly of the sequence. The chromosome 7 sequence positions of the identified STSs were

plotted relative to their established map positions (Fig. 2). Less than 1% of the identified STSs in the YAC-based map are in serious disagreement (defined as >3 megabases (Mb)) with their sequence position. In only one and four instances of the Genethon genetic and RH maps, respectively, are there discrepancies of more than several megabases between the map and sequence position of the STSs.

Together, these findings reveal an excellent overall concordance between the chromosome 7 sequence and previously constructed physical and genetic maps. The rare discrepancies can be largely accounted for by the inherent lower-resolution nature of the various mapping methods; however, it is also possible that some of the observed differences reflect polymorphisms between the different copies of chromosome 7 used for map construction and sequence generation. Nonetheless, these results, in conjunction with the robustness of the bacterial artificial chromosome (BAC)-based physical map used for sequence generation³, provide strong support for the established chromosome 7 sequence.

Orthology to mouse

The relationship between the human chromosome 7 sequence and the mouse genome could be readily defined for approximately 92%

Figure 1 Overview of chromosome 7. Headings at each end of the figure indicate the following (from top to bottom): Cytogenetic map; Gaps, positions of gaps within the chromosome 7 sequence; Orthologous mouse, orthologous mouse regions; (G + C) content, (G + C) content (blue) in 20-kb windows (scale = 30–65%); Repeat density, SINES (red) and LINES (blue) in 50-kb windows (scale = 0–70%); Interchromosomal, duplicated interchromosomal regions (yellow); Intrachromosomal, duplicated intrachromosomal regions (red); Genethon markers (green); CpG islands (blue), 200-bp windows on repeat-masked sequence; RNA genes, non-coding RNAs (ncRNA genes in orange; ncRNA predicted pseudogenes in brown); Pseudogenes, 573 processed (purple), 81 non-processed (lime) and 287 unknown (light purple) pseudogenes; Known genes, known (blue) and predicted (red) genes and gene identifiers.

of the chromosome, with 26 identifiable segments sharing the same order of highly conserved sequences in the two species at a resolution of 300 kb (ref. 9) or 46 segments at a resolution of 100 kb (Supplementary Methods). The smallest and largest defined segments are 200 kb and 38 megabases (Mb), respectively, with the latter residing on 7q and containing the cystic fibrosis gene, the ST7 tumour suppressor gene¹⁰, olfactory and taste receptors^{11,12}, and the T-cell receptor beta gene region¹³.

General features

We analysed the chromosome 7 sequence for interspersed repeat content, (G + C) content and the presence of CpG islands. The sequence has an overall repeat content (47%) and distribution of individual repeat classes (for example, short interspersed nucleotide elements (SINEs), long interspersed nucleotide elements (LINEs) and long terminal repeats (LTRs)) that differ only slightly from the whole-genome averages⁴ (Supplementary S1). Similarly, its overall (G + C) content (41%) is almost identical to that of the genome as a whole. The regions of highest (G + C) content flank the remaining gaps, consistent with the idea that the (G + C)-rich regions may be difficult to clone. Analysis of the repeat-masked chromosome 7 sequence revealed 1,461 CpG islands. Of the known chromosome-7-derived messenger RNAs (see below), the 5' end of 66% were at or near (5-kb upstream to 1-kb downstream of) a CpG island. For the full gene set presented here, the number with overlapping CpG islands was 51%—the two values bracketing the reported figure of 60% for the genome as a whole¹⁴. Table 2 provides a comparison of some of the general features of chromosome 7 with the other published chromosomes.

Previous analyses of the draft human genome sequence suggested that Alu repeat distribution correlated more strongly with gene content than with (G + C) content⁴. We re-examined this issue by correlating (G + C) content, Alu repeat content and known exonic sequences (see below) across chromosome 7. Using a range (50–800 kb) of non-overlapping windows, we found strong positive correlations between (G + C) content, exonic sequence density and Alu content. However, for all window sizes, Alu elements were more strongly correlated with (G + C) content than with exonic sequence density ($R^2 = 0.66$ compared with $R^2 = 0.41$ for 200-kb windows).

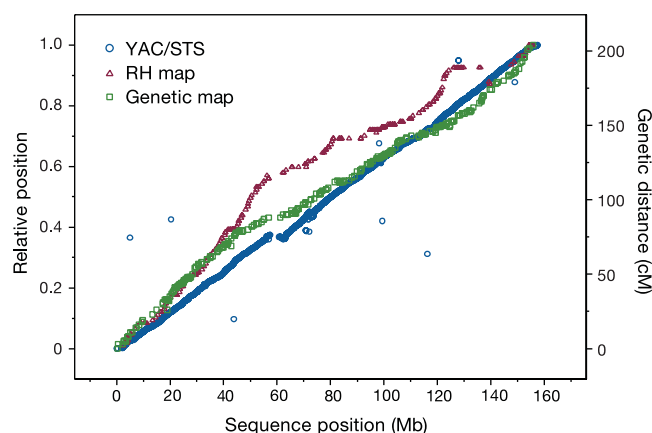


Figure 2 Comparisons of mapped positions of STSs and their locations within the chromosome 7 sequence. The position of each identified STS in the chromosome 7 sequence is plotted relative to its position on the chromosome 7 YAC-based map (blue), Genethon genetic map (green), and RH map (red). For the YAC/STS and RH maps, the left y axis reflects relative STS positions for those maps. For the Genethon genetic map, the right y axis reflects the actual genetic distances in cumulative centimorgans (cM). The apparent break at approximately 57–60 Mb reflects the position of the centromere.

This difference from the overall genomic pattern suggests that the factors influencing Alu distribution are complex.

Known protein-coding genes

The identification of genes within genomic sequence typically uses known mRNA sequences¹⁵, *ab initio* methods¹⁶ and comparative nucleotide or protein sequence data^{17,18}. Recognizing the challenges of gene annotation¹⁹ we sought to establish a foundation for the gene catalogue of human chromosome 7 as an intermediate step en route to a full understanding of all the functional elements encoded by the chromosome. First, a set of 1,073 human mRNAs from REFSEQ¹⁹ and the Mammalian Gene Collection²⁰ were uniquely assigned to the sequence and manually edited, resulting in 605 non-overlapping mRNAs with 45% showing alternative splicing. Only two (<0.3%) of the known genes mapped to chromosome 7 were not identified in the existing sequence.

Detailed examination of these mRNAs aligned against the genome revealed some potential artefacts, even within this experimentally supported set. For example, 23 mRNAs (2.4%) had no similarities to any mouse gene or to any known protein in the database. By contrast, less than 1% of predicted mouse genes have no homologue in the human genome⁹. Although some of these could be true genes, they may also represent untranslated segments of bona fide genes or transcripts of uncertain function. Nonetheless, these were retained in the current set. Additionally, eight other mRNAs had no significant open reading frame and were not included in the final gene set.

We also investigated 61 mRNAs where the matched genomic sequence contained differences from the mRNA that caused in some cases a frameshift and/or truncation of the protein product. To determine the origin of the difference, we re-sequenced the region of interest in a panel of 24 diverse individuals²¹, in the starting BAC and in some cases in overlapping BACs. Ten cases could not be uniquely amplified because of surrounding repetitive sequence. In six cases the BAC sequence was found to be in error (representing either a simple sequence error or a mutation at the site during propagation), and the sequence was corrected. For another 35 cases, there was support only for the BAC sequence. Using expressed sequence tag (EST) data, underlying mRNA data and conservation with mouse, these 35 cases were determined to be mRNA error (primarily a deletion or insertion with a second compensatory insertion/deletion downstream). In the remaining ten cases, the site was found to be polymorphic, with support for both BAC and mRNA sequence. One of these polymorphisms is a deletion in zonadhesin (ZAN)²², a sperm membrane protein that binds in a species-specific manner to the extracellular matrix (zona pellucida) of the egg. This deletion creates a frameshift at position 1,922 of this 2,812 amino acid protein.

Predicted protein-coding genes

To predict additional genes on chromosome 7, we applied three gene prediction programs. One, Genewise¹⁷, uses protein homologies to

Table 2 Comparison of human genome and published chromosomes

Category*	Chromosome					Human genome
	7	14	20	21	22	
(G + C) content (%)	41.0	40.9	44.1	40.9	47.8	41.0
Repetitive content (%)	45.0	46.2	42.0	40.1	41.9	44.8
Gene coverage (%)	36.5	43.6	42.4	31.0	50.0	27.0
Exon coverage (%)	1.4	2.3	2.4	—	5.0	1.5
Gene density (per Mb)	7.5	10.0	12.2	6.7	16.3	~10.0
Known gene mean size (kb)	61.5	58.7	51.3	57.0	—	27.0
All genes mean size (kb)	45.8	45.7	27.6	36.0	34.1	—
Pseudogenes (% of total genes)	45.0	26.0	18.8	20.8	25.0	—
Sequence length (Mb)	153.8	87.4	59.2	33.5	33.5	2,737.0

*Differences in gene features between chromosomes may in part reflect methodological differences: chromosome 14 (ref. 44), chromosome 20 (ref. 45), chromosome 21 (ref. 46), chromosome 22 (ref. 49), genome (ref. 19).

seed prediction, and the other two, Twinscan^{18,23} and FGENESH2¹⁶, use comparative sequence analysis. We used all available protein predictions²⁴ for Genewise and the mouse genome sequence⁹ as informant sequence for Twinscan and FGENESH2. The combined output predicted 90% of all known exons and 98% of the known genes, indicating that for known genes the combined output is reasonably comprehensive with high sensitivity, albeit at the cost of specificity.

To reduce the number of false positives and pseudogenes in the collection, we demanded that the predicted genes have a highly significant match in the mouse gene set in the orthologous region of mouse where possible, and in turn that the matching mouse gene have among its best matches the original chromosome 7 predicted gene ('reciprocal match'). Furthermore, single-exon genes were removed from the collection if they had matches to multi-exon genes in either the human or mouse genomes. Redundancy between the three sets among themselves and with known genes was eliminated, accepting in order known genes, FGENESH2, Twinscan and Genewise predictions, with gene models with a reciprocal match taking precedence. Predictions showing signs of non-functionality (truncation or absence of introns) and those that produced L1/reverse transcriptase were also removed from the set. This yielded 545 predicted genes, bringing the total number of protein-coding genes on chromosome 7 to 1,150.

We next examined ESTs to look for genes that the above process may have missed. Of 41,399 spliced ESTs that had their best match to chromosome 7, 93% at least partially overlapped existing exons, and an additional 1% lay near or within existing genes and suggested alternative splice forms (all were represented in the redundant gene prediction set). The remainder lacked significant open reading frames, and none satisfied the reciprocal match criteria used for the gene predictions above (indeed only 5% had any match to mouse predicted genes). Although these unplaced spliced ESTs could represent protein-coding genes, there is currently little corroborative evidence that they do. More likely they represent other transcription products, including non-coding RNA genes or untranslated fragments of protein-coding genes.

By several criteria, the predicted gene set is robust. As expected given the methods used to establish it, 94% have a reciprocal best match (that is, the predicted gene has as its best match a mouse gene, which in turn has as its best match the starting human gene) with the mouse genome (known genes show a 92% best match). The remainder have a reciprocal match. High percentages of the predictions are supported by similarities to non-mammalian vertebrate genomes and by EST matches (Table 3a), but as expected the rates for each of these are not as high as seen for known genes. The known

genes are enriched in highly expressed genes, which are more likely to be represented in EST sets and are more likely to be conserved across evolution. The predicted genes also compare favourably to known genes in coding exon number and in total coding sequence (Table 3b). The fact that both exon number and total coding sequence is smaller for the predicted genes suggests that either some terminal exons have been missed or that there are fragmented genes in the set that reduce average values. Finally, the pseudogene analysis carried out below shows that the set is remarkably free of likely pseudogenes, and that the set is not missing many genes (<60) with similarities to known proteins.

Pseudogenes

To identify pseudogenes on chromosome 7, we adapted an approach used for analysing the *Anopheles*²⁵ and mouse⁹ genome sequences. This involved identifying sequence with significant similarity to known proteins in regions that reside between known or predicted genes, and then evaluating the ratio of non-synonymous to synonymous coding changes (the K_A/K_S ratio) for each potential coding sequence. Although not absolute, this ratio is an indicator of selective constraints associated with particular DNA regions, and can be used to assess differences between genes that evolve under purifying selection and pseudogenes that evolve in a neutral fashion²⁶. Of the 941 such regions identified, nearly all ($97 \pm 3\%$) seem to evolve under neutrality and therefore are considered to contain pseudogenes. In contrast, only $5 \pm 3\%$ of the predicted and known genes have a K_A/K_S ratio consistent with neutral evolution. As the search did not attempt to identify highly diverged copies and may have merged two or more pseudogenes in the same interval, this probably corresponds to a lower limit for the number of pseudogenes. Indeed, there may be more pseudogenes than true genes on chromosome 7. As with the mouse genome sequence⁹, a significant fraction (33%) of the identified pseudogenes contain neither stop codons nor frameshifts. Virtually all pseudogenes (94%) could be aligned to another region in the human genome with higher sequence identity than to any region in the mouse genome, suggesting that they originated after the human-mouse divergence.

Pseudogenes are generally formed by two independent mechanisms of duplication: retrotransposition (giving rise to processed pseudogenes) and segmental duplication (often leading to non-processed pseudogenes). In an attempt to classify the pseudogenes on chromosome 7, we exploited the available mouse sequence. First, in contrast to non-processed pseudogenes, processed pseudogenes integrate throughout the genome and are unlikely to have sequence similarity in orthologous mouse regions. Furthermore, pseudogenes that arose before the divergence of mouse and human are probably so diverged as to be below the thresholds of detection used here. For 654 pseudogenes within regions of chromosome 7 with identified mouse orthology, 573 (88%) appear to be processed, and 81 (12%) appear to be segmentally duplicated pseudogenes. Processed pseudogenes are broadly distributed along chromosome 7, with a slight tendency to cluster near the telomeres, whereas non-processed pseudogenes are concentrated in gene-rich regions (see Fig. 1). A minority of the former (100) has inserted into the introns of unrelated functional genes. The direction of these processed pseudogenes relative to the host intron varies almost randomly: 41 (41%) are integrated in the same strand as the host gene, whereas 59 (59%) are in the opposite strand.

Non-coding RNAs

We identified non-coding RNA (ncRNA) genes in the chromosome 7 sequence as described for the draft human genome sequence¹⁹, followed by a refinement step using matches to mouse orthologous regions. *De novo* computational gene-finding methods for most non-coding RNA genes are not yet sufficiently robust for automated genome annotation²⁷. Annotation was restricted to transfer RNA

Table 3 Coverage and characteristics of exons and genes on chromosome 7

a Coverage of predicted exons and genes by various data sets

	Predicted (%)		Known (%)		Ratio	
	Exons	Genes	Exons	Genes	Exons	Genes
Human ESTs	42	59	85	94	0.49	0.63
Non-mammalian*	20	38	36	52	0.56	0.73
Total	47	66	89	95	0.53	0.69
Pfam	—	45	—	69	—	0.65
Interpro	—	60	—	78	—	0.77

b Characteristics of predicted versus known genes

	Predicted	Known	Ratio
Exons per gene	6.7	9.5	0.71
Coding bases per gene (bp)	1,231	1,457	0.84
Genic bases per gene (bp)	28,474	61,439	0.44

* Translated non-mammalian vertebrate genomic sequence: *Fugu rubripes*, *Tetraodon nigroviridis*, *Gallus gallus*. Ratio is that of predicted to known.

genes, for which there is a robust identification programme, tRNAscan-SE²⁸, and to strong primary sequence similarities to known mammalian ncRNAs, including recently discovered human microRNAs²⁹.

Twenty-three tRNA genes were identified in the chromosome 7 sequence, including a cluster of 20 tRNAs with 18 tRNACys-GCA genes in one 400-kb interval. Eleven other ncRNA genes were found: four microRNAs, two U6 small nuclear RNA (snRNA) genes, and five genes for the four cytoplasmic Y RNAs (hY1, hY4, hY5 and two near-identical copies of hY3); all of the human Y RNA genes were already known to be located on chromosome 7 (ref. 30). Additionally, 302 putative pseudogenes were detected by BLAST similarity to other human ncRNA genes; this included 65 apparent U6 snRNA pseudogenes, 43 hY3 RNA pseudogenes and 37 non-Alu SRP-RNA pseudogenes.

Protein index

Using the above-generated gene set, we derived an index of predicted protein sequences. In turn, this was compared to the Interpro database³¹ using Interproscan³², which predicts protein families, domain and repeat families, and sequence motifs. The Interpro results were used to assign Gene Ontology (GO) codes³³: 51% of proteins were in the category of molecular function, 31% in biological process and 18% in cellular components. Of the 68% of the proteins that had an Interpro classification, 56% were multi-domain. The two most prevalent families are also two of the most prevalent in the human genome: the immunoglobulins and zinc fingers.

The general homeobox domain proteins are heavily represented on chromosome 7, accounting for 21 of the 211 currently annotated records in the Ensembl gene set³⁴ of the human genome. In addition, the chromosome 7 gene set contains one-third of each of the more

specific homeobox proteins in the genome: antennapedia homeobox proteins, engrailed-type homeobox proteins and homeo-domain protein CUT domains. The chromosome 7 HOX region, one of the four homeobox clusters in the human genome, contains ten two-exon genes (with one alternative form). This 90.7-kb region has a (G + C) content of 52% and is completely devoid of known interspersed repeats.

Williams syndrome critical region

Williams–Beuren syndrome (WBS) is associated with large (typically about 1.6 Mb) deletions³⁵ within 7q11.23. A series of large (>140 kb) duplicated segments (duplicons) span over 2 Mb in this region and are shown as coloured boxes in Fig. 3. The commonly deleted region is flanked by a duplicon that contains the p47-phox gene (or pseudogene), which has served as a useful marker for differentiating among the different duplicated segments. A third copy of the duplicon resides distal to the other pair within the commonly deleted region, in inverted orientation. At present there is a gap in the physical map and the chromosome sequence between the second and third copies of this duplicon.

The WBS region presented the single greatest challenge to the mapping and sequencing of the euchromatic regions of chromosome 7. Indeed, even after considerable effort, there remains some uncertainty about the location and orientation of some sequences. The duplicated segments approach the size of BACs, and the frequency of sequence differences among duplicons approaches (or is less than) the human polymorphism rate. To add to this complexity, the long-range organization of the region, perhaps including the number of duplicons, differs among individuals.

In attempting to establish a representative sequence of the WBS region, we used clones from a single BAC library to reduce the allelic complexity to just two variants. We also deliberately sequenced

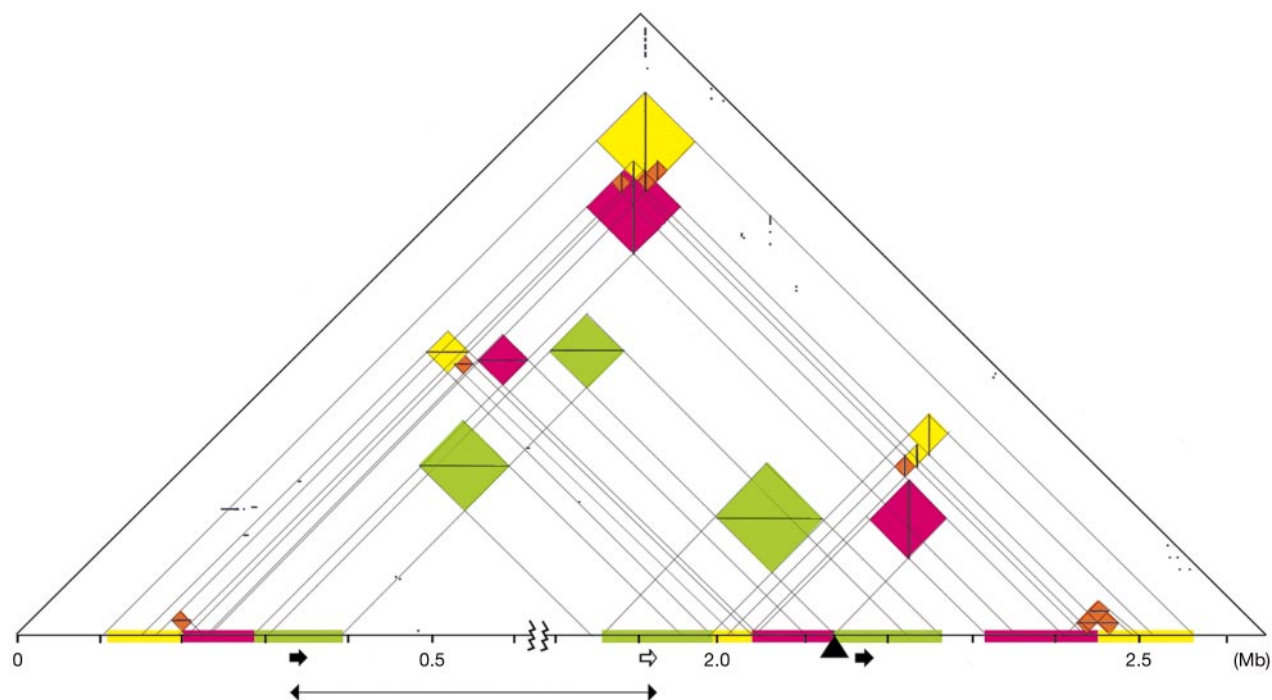


Figure 3 Repetitive content surrounding the commonly deleted region of the Williams–Beuren Syndrome area of 7q11.23 illustrated using self_dot_plot (H. Skaletsky and S. Rozen, personal communication). Each dot represents an identical match of 200 bp. Direct repeats are shown by horizontal lines, inverted repeats as vertical lines and palindromes as vertical lines that nearly intersect the baseline. Large sequence

duplications are indicated by coloured boxes. As indicated by the line break, a central 1.15 Mb of non-duplicated sequence has been omitted. The start of the region shown corresponds to base 70,945,566. Arrows mark the positions of the two phox genes (closed) and the phox pseudogene (open).

BACs with extensive overlaps to establish linkage of variant sites and used polymerase chain reaction (PCR) analysis of clones and a human population panel to distinguish polymorphic sites from differences between duplicated segments. This process eventually yielded the two sequence contigs shown in Fig. 3.

Segmental duplications

Segmental duplications are large low-copy repeats that arise as a consequence of duplication of genomic DNA and may range up to hundreds of kilobases in length. We performed a detailed analysis of duplicated sequence ($\geq 90\%$ sequence identity and ≥ 1 kb in length; see Methods), comparing the chromosome 7 sequence against a recent assembly of the human genome. We identified a total of 3,215 pairwise alignments that met these criteria of length and sequence identity (Fig. 4a; see also Supplementary S3 and S4), making chromosome 7 one of the most duplicated human chromosomes. Overall, 8.2% (12,588 kb) of the sequence shares sequence homology to more than one location in the genome (Fig. 4a; see also

Supplementary S4). The enrichment is predominantly due to an increase in intrachromosomal duplications (7.0% of the sequence) rather than interchromosomal duplications (2.2%) with 0.5% (729,982 bp) sequence overlap between the two types. The spatial distribution of the interchromosomal and intrachromosomal duplications is clustered (Fig. 4a; see also Supplementary S3 and S4). As expected, large blocks of interchromosomal duplication locate preferentially within the pericentromeric and subtelomeric regions^{36,37}. A marked asymmetry, however, was observed between the short and long arms of chromosome 7. The short arm of chromosome 7 has large blocks of recent interchromosomal duplications within both the pericentromeric (600 kb) and subtelomeric (about 150 kb) regions. The p arm subtelomere contains the most recent interchromosomal duplications ($>99\%$). The low degree of sequence divergence suggests that such regions may have duplicated and/or undergone gene conversion since the separation of the human and chimpanzee lineages from a common ancestor.

In contrast to the short arm, the q arm pericentromeric region shows a much smaller block of duplication (<200 kb) followed by a large (about 600 kb) tract of monomeric alpha-satellite repeat sequence. These duplications are highly divergent ($<93\%$). Although possibly a much more extensive domain of pericentromeric duplication remains to be sequenced closer to the centromere, the current configuration and the low degree of sequence similarity among these interchromosomal duplications suggest that the 7q pericentromeric region has been relatively quiescent over the last 25 million years of evolution. The subtelomeric region of chromosome 7q is even more striking in this regard. There is virtually no evidence of segmental duplication within a megabase of the telomere. The molecular mechanisms that underlie the difference in subtelomeric and pericentromeric duplication architecture between the two arms of chromosome 7 are unknown, but the asymmetry may represent a general property of metacentric chromosomes as has been suggested previously^{36,38}.

Comparison of the features of inter- and intrachromosomal duplications reveals some intriguing differences. In general, the intrachromosomal alignments tend to be larger than interchromosomal alignments (8 out of 10 pairwise alignments in excess of 100 kb were intrachromosomal duplications; Supplementary S5). This trend may, in part, be due to the fragmented nature of the draft sequence for other unfinished chromosomes. Furthermore, intrachromosomal duplications show an abundance of highly similar duplications ($>98\%$ identity), whereas most (57.1%) of the interchromosomally duplicated bases cluster between 93% and 96% identity (mode of 94.5–95.0%) (Fig. 4b; see also Supplementary S5). This interchromosomal mode is significantly different from the

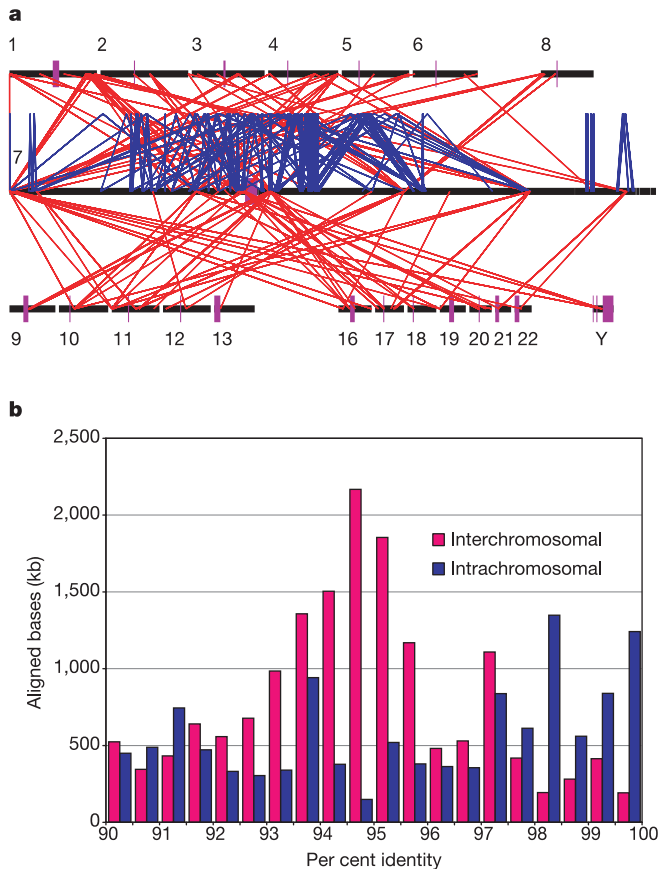


Figure 4 Recent segmental duplications on chromosome 7. **a**, Large (>10 kb) highly similar ($>95\%$) intrachromosomal (blue) and interchromosomal (red) segmental duplications are shown for chromosome 7. Chromosome 7 is magnified in scale relative to the other chromosomes. The intrachromosomal duplications shown may be sites of unequal homologous recombination resulting in large-scale rearrangements such as those that cause WBS. The centromere is coloured purple. For more details regarding these regions see Supplementary S3–S6. **b**, Sequence similarity of segmental duplications. For all pairwise alignments, the total number of aligned bases was calculated and binned based on per cent sequence identity. Sequence identity distributions for interchromosomally (red) and intrachromosomally (blue) duplicated bases are shown. Assuming a clock-like rate for nuclear DNA substitution, per cent similarity is linear with evolutionary time providing a surrogate for the age of the duplication or gene conversion event.

Table 4 Duplicated features for non-redundant transcripts					
Non-redundant source	Duplicated features		All features		
	Exons (%)	Transcripts (%)	Exons	Transcripts	
mRNA-based transcripts	533 (8.7%)	121 (20.6%)	6,122	588	
Gene	508 (8.6%)	117 (20.4%)	5,926	573	
Pseudogene	25 (12.8%)	4 (26.7%)	196	15	
Predicted transcripts	694 (22.8%)	136 (36.2%)	3,045	376	
Gene	353 (15.3%)	64 (26.0%)	2,314	246	
Pseudogene	341 (46.6%)	72 (55.4%)	731	130	
Spliced EST clusters	202 (19.3%)	91 (24.6%)	1,049	370	
Total	1,429 (14.0%)	348 (26.1%)	10,216	1,334	

The non-redundant transcript set was screened for confirmatory transcriptional support by best genomic placement of ESTs. mRNAs required one supporting EST, and EST clusters and predicted transcripts required two ESTs and evidence of splicing (two or more exons). Thus, EST clusters of a single EST were removed. Additionally, each gene feature was binned as duplicated if at least 50 bp overlapped a duplicated region. Thus, exons less than 50 bp were lost from this analysis. Overall, 14.0% of all exons were binned as duplicated. Similar results were obtained requiring the entire exon to be encompassed by duplication (13.8% of all exons duplicated). Coding potential was only assessed as part of the analysis of mRNA-based and predicted transcripts. For these, the pseudogene fraction of duplicated exons ($366/1,227 = 30\%$) was greater than for unique exons ($927/7,940 = 11\%$).

average computed for the draft genome as well as other published chromosomes (96.0–98.0%)³⁹. Using sequence divergence as an indicator of evolutionary age, the data suggest that chromosome 7 has been the target of more ancient interchromosomal duplication and/or gene conversion events.⁴⁰ In contrast, more recent chromosome-specific duplication/gene conversion events⁴⁰ have occurred since the separation of the human and African ape lineages.

Segmental duplications are known sites of both pathological and evolutionary instability^{41,42}. To identify regions on chromosome 7 that may be associated with genomic disorders, we searched for regions (between 50 kb and 10 Mb apart) that were flanked by large (≥ 10 kb), highly homologous ($\geq 95\%$) segmental duplications (Supplementary S6). In addition to the WBS region, we detected ten regions of chromosome 7 (Supplementary S7 and S8), nine of which contained genes and would be considered candidates for genomic disorders. In total, these 11 regions (corresponding to 103 pairwise alignments) of chromosome 7 implicate almost a quarter (40.7 Mb) of the chromosome as being susceptible to duplication-mediated rearrangement. It will be important to investigate these regions experimentally for large-scale variation and association with disease.

Segmental duplications have long been noted for their potential role in the evolution of new genes⁴³. To examine the transcriptional and coding potential of duplicated regions, we analysed a hierarchical, non-overlapping set of known genes, predicted genes and remaining spliced EST clusters (see above). For each group, we categorized every exon as unique or duplicated on the basis of its overlap with duplicated sequence (Table 4). Almost 14% (1,244 out of 9,890) of all exons are duplicated, and most of these lie within intrachromosomal duplicated sequence as opposed to interchromosomal duplicons. Our analysis shows that the relative number of transcribed exons is significantly greater for duplicated DNA when compared with non-duplicated DNA on chromosome 7. These results support a previous observation³⁹ that recently duplicated regions are rich in genes/transcripts. It should be noted, however, that many transcripts within the duplicated sequence have poor translational potential (44% compared with 20% of exons in pseudogenes for duplicated and unique regions), based on analysis of open reading frames (see pseudogene analysis above). Indeed, our analyses suggest that genes within duplicated regions show relaxed selective constraint when compared with genes encoded within unique portions of chromosome 7. Duplicate regions of chromosome 7 are, therefore, enriched for a particular class of pseudogene, which may be transcribed and may possess intron–exon structure but is unlikely to be translated. Most of the pseudogenes probably represent dying transcripts, which may on rare occasions lead to the formation of new genes.

Conclusions

The sequence of human chromosome 7 described here, and that of several other human chromosomes^{44–47}, represent landmark steps in the Human Genome Project. As chromosome sequences advance from their initial 'draft' status to a high-accuracy comprehensive stage, the molecular landscape becomes clearer, and the ability to perform detailed analyses becomes more robust. For our studies of chromosome 7, the combination of a high-quality, nearly complete sequence and a draft sequence of the mouse genome allowed us to perform rigorous gene analyses that included an improved ability to distinguish pseudogenes from bona fide genes. In addition to generating a gene index for the chromosome, our data provide evidence for a small subset of proteins that contain a polymorphism leading to a truncated protein in the human lineage. A chromosome-wide view of segmental duplications revealed that, compared with other chromosomes analysed so far, chromosome 7 exhibits a much higher rate of intrachromosomal duplication. Furthermore, there seems to be evolutionary asymmetry between the long and short arms. Taken together, these findings illustrate the dynamic

nature of a mammalian chromosome. Such dynamic behaviour also has adverse consequences, as revealed by the sequence features of the WBS region, where large duplicons of remarkably high sequence similarity mediate disease-causing deletions. Finally, the sequence we report here for chromosome 7 has directly facilitated the identification of a number of genes associated with human disease (for example, refs 10, 48). But these examples, although highly gratifying, simply represent the beginning of efforts to capitalize on the knowledge provided by finished genomic sequence for better understanding the genetic bases for human health and disease. □

Methods

Tiling-path verification

To evaluate clone overlaps where the rate of difference between overlapping clone sequences was higher than 1 in 1,000 bases, a PCR product encompassing the differences was sequenced from each BAC in the overlap region and from each of a panel of 24 ethnically diverse genomic DNA samples²¹. If the 24 samples showed allelic variation, the overlap was judged to be correct, but if the 24 samples yielded persistent heterozygosity, the sequence was judged to be derived from a repeated sequence, with sequence differences between the copies.

Assaying mRNA/genomic discrepancies

To investigate discrepancies between mRNA and genomic sequence, PCR products were generated, re-sequenced and the polymorphic bases examined in the DNA from 24 individuals, the original BAC, and in some cases other BACs.

Pseudogene detection

After masking all predicted and known genes and common repeats, we performed homology searches by comparing the DNA sequence of chromosome 7 with a non-redundant protein database (see Supplementary Information). All regions matching non-viral and non-transposon known proteins (E value < 0.001) were further processed by merging those likely to be parts of the same gene or pseudogene. This step, although likely to cause the loss of some real pseudogenes, is essential to eliminate possible fragmented predictions, which could otherwise lead to an overestimation. We refined the prediction for the resulting DNA fragments by comparing them with the closest protein sequence using Genewise¹⁷. We finally confirmed the integrity of the predictions by removing elements without significant matches (E values < 0.001) in a second round of BLASTX against NRDB.

Each candidate pseudogene and all predicted and known genes were subjected to a K_A/K_S analysis. We first inferred the ancestral sequence of each of the target sequences (A) using a protein-based DNA multiple alignment of A and its two closest matches in NRDB between 50% and 95% identical. We next estimated the number of synonymous and non-synonymous substitutions occurring in sequence A by comparing it with its ancestral sequence using the YN00 program. In addition to the program's pre-set, we excluded those K_A/K_S ratios that were based on excessively low (< 20) or high ($K_S > 1$) numbers of substitutions, resulting in reliable calculations for approximately 50% of analysed sequences.

The fraction of neutrally evolving sequences included in the pseudogene and gene sets was calculated comparing their K_A/K_S distributions with benchmark distributions for functional and pseudogenic elements. These distributions were obtained from the K_A/K_S analysis of 2,000 functional human genes randomly selected from a 50% non-redundant RefSeq¹⁹ (reviewed) collection, and 1,730 processed pseudogenes with open reading frame truncations identified from a homology search through the whole human genome as performed for chromosome 7 (manuscript in preparation). Each of the K_A/K_S distributions associated to the pseudogene and gene sets of chromosome 7 was compared with benchmark distributions using a 'least squares fitting' to obtain estimates of neutrally evolving fractions of sequences. The error rate associated with this estimate is $< 3\%$ according to cross-validation analysis with the benchmark sequences.

We compared each identified pseudogene at the protein level with all mouse genes (Ensembl¹⁴) located in the corresponding orthologous regions. Positive match was considered when the associated E -value was $< 10^{-8}$. To avoid mis-annotations owing to the possible absence of decisive sequences in the mouse gene set, we also compared each translated pseudogene with the whole mouse orthologous region using tBLASTn with the same E -value cutoff as above.

Received 25 February; accepted 23 April 2003; doi:10.1038/nature01782.

1. Riordan, J. R. *et al.* Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245**, 1066–1073 (1989).
2. Liang, H. *et al.* Molecular anatomy of chromosome 7q deletions in myeloid neoplasms: evidence for multiple critical loci. *Proc. Natl Acad. Sci. USA* **95**, 3781–3785 (1998).
3. Human Genome Mapping Consortium. A physical map of the human genome. *Nature* **409**, 934–941 (2001).
4. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
5. Riethman, H. *et al.* Integration of telomere sequences with the draft human genome sequence. *Nature* **409**, 948–951 (2001).
6. Felsenfeld, A., Peterson, J., Schloss, J. & Guyer, M. Assessing the quality of the DNA sequence from the Human Genome Project. *Genome Res.* **9**, 1–4 (1999).

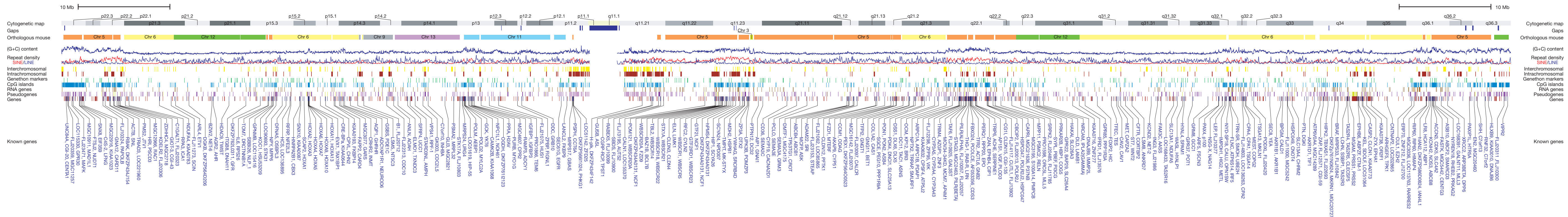
7. Bouffard, G. G. *et al.* A physical map of human chromosome 7: an integrated YAC contig map with average STS spacing of 79 kb. *Genome Res.* **7**, 673–692 (1997).
8. Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**, 152–154 (1996).
9. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
10. Zenklusen, J. C., Conti, C. J. & Green, E. D. Mutational and functional analyses reveal that ST7 is a highly conserved tumor-suppressor gene on human chromosome 7q31. *Nature Genet.* **27**, 392–398 (2001).
11. Nadezhdin, E. V. *et al.* Identification of paralogous HERV-K LTRs on human chromosomes 3, 4, 7 and 11 in regions containing clusters of olfactory receptor genes. *Mol. Genet. Genom.* **265**, 820–825 (2001).
12. Chandrasekar, J. *et al.* T2Rs function as bitter taste receptors. *Cell* **100**, 703–711 (2000).
13. Rowen, L., Koop, B. F. & Hood, L. The complete 685-kilobase DNA sequence of the human beta T cell receptor locus. *Science* **272**, 1755–1762 (1996).
14. Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc. Natl Acad. Sci. USA* **90**, 11995–11999 (1993).
15. Wheelan, S. J., Church, D. M. & Ostell, J. M. Spidey: a tool for mRNA-to-genomic alignments. *Genome Res.* **11**, 1952–1957 (2001).
16. Solovyev, V. V. *Statistical Approaches in Eukaryotic Gene Prediction* (ed. Balding, D.) (Wiley, New York, 2001).
17. Birney, E. & Durbin, R. Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* **10**, 547–558 (2000).
18. Korf, I., Flicek, P., Duan, D. & Brent, M. R. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17**, 140–148 (2001).
19. Pruitt, K. D. & Maglott, D. R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29**, 137–140 (2001).
20. Mammalian Gene Collection Program Team Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl Acad. Sci. USA* **99**, 16899–16903 (2002).
21. The International Human SNP Working Group A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
22. Hardy, D. M. & Garbers, D. L. A sperm membrane protein that binds in a species-specific manner to the egg extracellular matrix is homologous to von Willebrand factor. *J. Biol. Chem.* **270**, 26025–26028 (1995).
23. Flicek, P., Keibler, E., Hu, P., Korf, I. & Brent, M. R. Leveraging the mouse genome for gene prediction in human: From whole-genome shotgun reads to a global synteny map. *Genome Res.* **13**, 46–54 (2003).
24. Benson, D. A. *et al.* GenBankM. *Nucleic Acids Res.* **30**, 17–20 (2002).
25. Zdobnov, E. M. *et al.* Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* **298**, 149–159 (2002).
26. Ohta, T. & Ina, Y. Variation in synonymous substitution rates among mammalian genes and the correlation between synonymous and non-synonymous divergences. *Mol. Evol.* **41**, 717–720 (1995).
27. Eddy, S. R. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinform.* **2**, 18 (2002).
28. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
29. Mourelatos, Z. *et al.* miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes Dev.* **16**, 720–728 (2002).
30. Maraia, R. J., Sasaki-Tozawa, N., Driscoll, C. T., Green, E. D. & Darlington, G. J. The human Y4 small cytoplasmic RNA gene is controlled by upstream elements and resides on chromosome 7 with all other hY scRNA genes. *Nucleic Acids Res.* **22**, 3045–3052 (1994).
31. Mulder, N. J. *et al.* InterPro: an integrated documentation resource for protein families, domains and functional sites. *Brief Bioinform.* **3**, 225–235 (2002).
32. Zdobnov, E. M. & Apweiler, R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).
33. The Gene Ontology Consortium Creating the Gene Ontology resource: design and implementation. *Genome Res.* **11**, 1425–1433 (2001).
34. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res.* **30**, 38–41 (2002).
35. Korenberg, J. R. *et al.* Genome structure and cognitive map of Williams syndrome. *J. Cogn. Neurosci.* **12**, 89–107 (2000).
36. Horvath, J. E., Bailey, J. A., Locke, D. P. & Eichler, E. E. Lessons from the human genome: transitions between euchromatin and heterochromatin. *Hum. Mol. Genet.* **10**, 2215–2223 (2001).
37. Mefford, H. C. & Trask, B. J. The complex structure and dynamic evolution of human subtelomeres. *Nature Rev. Genet.* **3**, 91–102 (2002).
38. Eichler, E. E., Archidiacono, N. & Rocchi, M. CAGGG repeats and the pericentromeric duplication of the hominoid genome. *Genome Res.* **9**, 1048–1058 (1999).
39. Bailey, J. A. *et al.* Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. *Am. J. Hum. Genet.* **70**, 83–100 (2002).
40. DeSilva, U., Massa, H., Trask, B. J. & Green, E. D. Comparative mapping of the region of human chromosome 7 deleted in Williams syndrome. *Genome Res.* **9**, 428–436 (1999).
41. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
42. Stankiewicz, P. & Lupski, J. R. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* **18**, 74–82 (2002).
43. Ohno, S., Wolf, U. & Atkin, N. Evolution from fish to mammals by gene duplication. *Hereditas* **59**, 169–187 (1968).
44. Heilig, R. *et al.* The DNA sequence and analysis of human chromosome 14. *Nature* **421**, 601–607 (2003).
45. Deloukas, P. *et al.* The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414**, 865–871 (2001).
46. Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405**, 311–318 (2000).
47. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
48. Everett, L. A. *et al.* Pendred syndrome is caused by mutations in a putative sulphate transporter gene (PDS). *Nature Genet.* **17**, 411–422 (1997).
49. Collins, J. E. *et al.* Reevaluating human gene annotation: A second-generation analysis of chromosome 22. *Genome Res.* **13**, 27–36 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank the following laboratories, which contributed sequence segments to the final chromosome 7 sequence: Whitehead Institute Center for Genome Research; Institute of Molecular Biotechnology, Jena, Germany; Sanger Institute; University of Washington Multimegabase Sequencing Center; University of Oklahoma Advanced Center for Genome Technology; Stanford DNA Sequencing and Technology Development Center; Joint Genome Institute; Human Genome Studies Group, Los Alamos. We also acknowledge the efforts of the respective staff of the Washington University Genome Sequencing Center and the University of Washington Genome Center. We thank E. Birney for assistance in producing Fig. 1.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.K.W. (rwilson@watson.wustl.edu). Accession numbers for the sequence analysed for this paper can be found in Table 1. All reported DNA sequences have been deposited in GenBank or EMBL. The updated chromosome 7 sequence can be accessed through GenBank accession BL000002.



The recent expansion of Pluto's atmosphere

J. L. Elliot^{*†‡}, A. Ates[§], B. A. Babcock^{||}, A. S. Bosh^{¶¶}, M. W. Buie[‡], K. B. Clancy^{*}, E. W. Dunham[‡], S. S. Eikenberry[#], D. T. Hall[☆], S. D. Kern^{*}, S. K. Leggett[☆], S. E. Levine^{**}, D.-S. Moon[#], C. B. Olkin[‡], D. J. Osip^{*††}, J. M. Pasachoff^{‡‡}, B. E. Penprase[§], M. J. Person^{*}, S. Qu^{§§}, J. T. Rayner^{||||}, L. C. Roberts Jr[☆], C. V. Salyk^{*}, S. P. Souza^{‡‡}, R. C. Stone^{**}, B. W. Taylor[‡], D. J. Tholen^{¶¶}, J. E. Thomas-Osip^{*††}, D. R. Ticehurst^{‡‡} & L. H. Wasserman[‡]

^{*} Department of Earth, Atmospheric, and Planetary Sciences, [†] Department of Physics, ^{§§} Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139-4307, USA

[‡] Lowell Observatory, Flagstaff, Arizona 86001, USA

[§] Department of Physics & Astronomy, Pomona College, 610 N. College Avenue, Claremont, California 91711, USA

^{||} Physics Department, ^{‡‡} Hopkins Observatory, Williams College, Williamstown, Massachusetts 01267-2565, USA

^{¶¶} Institute for Astrophysical Research, Boston University, 725 Commonwealth Avenue, Boston, Massachusetts 02215, USA

[#] Department of Astronomy, 212 Space Sciences Building, Cornell University, Ithaca, New York 14853, USA

[☆] The Boeing Company, 535 Lipoa Pkwy, Suite 200, Kihei, Hawaii 96753, USA

^{**} US Naval Observatory, Flagstaff Station, PO Box 1149, Flagstaff, Arizona 86002-1149, USA

^{||||} NASA Infrared Telescope Facility, ^{¶¶} Institute for Astronomy, 2680 Woodlawn Drive, Honolulu, Hawaii 96822, USA

Stellar occultations—the passing of a relatively nearby body in front of a background star—can be used to probe the atmosphere of the closer body with a spatial resolution of a few kilometres (ref. 1). Such observations can yield the scale height, temperature profile, and other information about the structure of the occulting atmosphere. Occultation data acquired for Pluto's atmosphere in 1988 revealed a nearly isothermal atmosphere² above a radius of $\sim 1,215$ km. Below this level, the data could be interpreted as indicating either an extinction layer or the onset of a large thermal gradient, calling into question the fundamental structure of this atmosphere. Another question is to what extent Pluto's atmosphere might be collapsing as it recedes from the Sun (passing perihelion in 1989 in its 248-year orbital period), owing to the extreme sensitivity of the equilibrium surface pressure to the surface temperature. Here we report observations at a variety of visible and infrared wavelengths of an occultation of a star by Pluto in August 2002. These data reveal evidence for extinction in Pluto's atmosphere and show that it has indeed changed, having expanded rather than collapsed, since 1988.

Pluto's predominantly nitrogen atmosphere is in vapour pressure equilibrium with the surface ice, and consequently can undergo large changes in pressure in response to small changes in surface-ice temperature³. In 1988, when this atmosphere was probed with a stellar occultation from a variety of sites^{4–6}, the light curve revealed a pronounced 'kink' (an abrupt change of slope, described below) that could be interpreted as the abrupt onset of either a thermal gradient^{7,8} or an extinction layer^{2,5}. The kink was observed during the onset of occultation ('immersion', also known as 'ingress') and at its termination ('emersion', also known as 'egress'). In order to resolve these issues, attempts have been made for over a decade to observe additional Pluto occultations^{9,10}. None proved successful until our observations of the 20 July 2002 occultation of the star 'P126A'¹⁰, from which we determined that Pluto's atmosphere had changed^{11,12} since 1988, but the data set was not sufficient for further conclusions¹¹.

The 21 August 2002 occultation of the star 'P131.1'¹⁰ (R-band magnitude $R = 15.7$ mag, K-band magnitude $K = 13.3$ mag) was predicted (<http://occult.mit.edu/research/occultations/Candidates/Predictions/P131.1.html>), and subsequently observed from several telescopes in Arizona, California and Hawaii (see Fig. 1, Table 1, and Supplementary Table 1). Although most of these observations were at visible wavelengths with no filtering, we probed for possible extinction with additional wavelength coverage: an H-band infrared filter on the UK Infrared Telescope (UKIRT) and continuous coverage between 0.8 and 2.5 μm with the SpeX^{13,14} spectrograph on the Infrared Telescope Facility (IRTF). This data set is unique, in that (1) it provides wide spatial coverage over Pluto's disk, which is essential for the accurate astrometric solution needed for derivation of the atmospheric information, and (2) it has the wavelength coverage needed to allow some distinction between extinction and differential refraction effects. The data from the University of Hawaii 2.2-m telescope (UH 2.2-m) had the best signal-to-noise ratio and the highest time resolution—0.5 s, during which Pluto's shadow moved only 3.3 km (the highest spatial resolution yet achieved within Pluto's atmosphere). These data are plotted as the solid line in Fig. 2, along with the 1988 data from the Kuiper

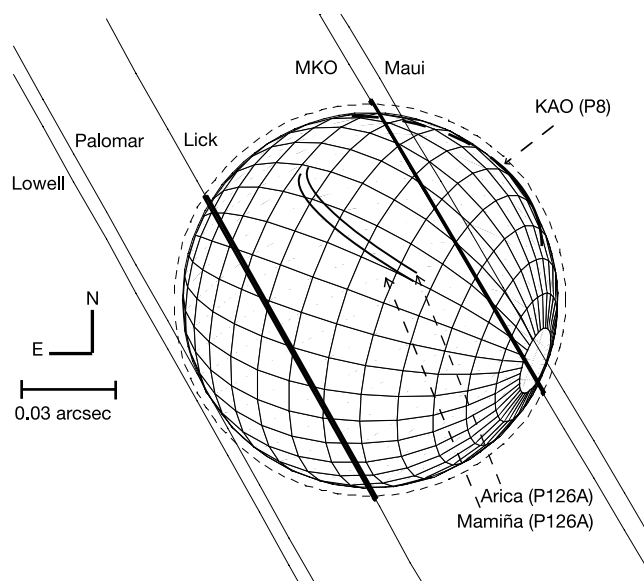


Figure 1 Occultation probes of Pluto's atmosphere. Occultation chords are overlaid on a schematic of Pluto, which is plotted as it appeared during the P131.1 occultation with the north rotational pole (IAU definition) behind the limb of the planet (sub-Earth latitude and east longitude of -28.1° and 38.1° , respectively). Straight lines indicate the occultation chord paths for all successful observing stations. The Lowell chord grazed the upper atmosphere with the stellar flux dropping only 6%. The two bold chords for Mauna Kea Observatory (MKO) and Lick represent the data used to calculate the astrometric solution upon which other analyses in this work was based. For this solution, the closest approach to the centre of Pluto's shadow was 597 ± 32 km for the MKO telescopes and 600 ± 32 km for Lick, and the velocities of Pluto's shadow were 6.8489 km s^{-1} and 6.7293 km s^{-1} for the two sites, respectively. The dotted circle indicates the solution's half-light radius (in the shadow) of $1,213 \pm 16$ km, which compares with $1,154 \pm 20$ km determined from the 1988 data. Assuming no extinction above the half-light levels, these correspond respectively to $1,283 \pm 9$ km and $1,214 \pm 20$ km in Pluto's atmosphere, after correcting for the smaller size of the shadow, due to refraction⁵. Over-plotted are the equivalent ground-paths of the P126A Mamiña chord obtained by our group¹¹ as well as the P126A Arica chord¹² and the P8 KAO chord⁵. These curves are the straight chord lines from the past events^{5,11,12} projected onto P131.1 Pluto coordinates. The solid (right) portion of the KAO chord is in front of the planet, while the dashed (left) portion is behind the planet's limb. All chord directions (immersion to emersion) are from southwest to northeast, except the KAO chord which is reversed.

^{*} Present address: Las Campanas Observatory, Carnegie Observatories, Casilla 601, La Serena, Chile.

Table 1 Observations from Mauna Kea and Lick observatories

Site, telescope (aperture, in m)	East longitude (° ' ")	Latitude (° ' ")	Instrument	Wavelength	Recording interval (21 Aug. 2002 UT)	Cycle time (s)	Integration time (s)	SNR*	Observers
Lick									
Shane (3.0)	−121 38 14	37 20 34	PCCD ²³	Visible†	06:34:00–07:03:58	1.0	1.0	16	E.W.D., C.B.O.
Mauna Kea‡									
IRTF (3.0)	−155 28 19	19 49 34	SpeX ^{13,14}	0.8–2.5 µm	06:29:09–07:11:20	12.7–15.5§	10.0	57	J.L.E., K.B.C., J.T.R.
UH (0.6)	−155 28 16	19 49 18	PCCD ^{23,24}	Visible†	06:34:26–07:11:25	7.0–8.0§	5.0	24	M.J.P., S.Q.
UH (2.2)	−155 28 10	19 49 23	Williams CCD	Visible†	06:39:49–06:59:49	0.5	0.5	103	J.M.P., B.A.B., D.R.T.
UKIRT (3.8)	−155 28 13	19 49 21	IRCAM ²⁵	H	06:41:12–06:57:03	3.3–4.8§	1.5	39	D.J.T., D.J.O., S.K.L.

For a complete listing of all attempted observations, see Supplementary Table 1 and the independent results in the companion paper¹².

* 'SNR' is the signal-to-noise ratio for the signal from the unocculted star integrated over a time interval corresponding to 60 km (about a scale height) of shadow motion¹⁷. The SNR for the Lick data was lower than expected from the size of the telescope aperture compared with other sites due to poor observing conditions (5-arcsec images at Lick compared with 0.5-arcsec images at Mauna Kea Observatory).

† 'Visible' means that an unfiltered CCD was used.

‡ For the Mauna Kea Observatory telescopes: IRTF is the Infrared Telescope Facility, UH is the University of Hawaii, and UKIRT is the United Kingdom Infrared Telescope.

§ Variable cycle times occur for a variety of reasons that will be discussed elsewhere.

Airborne Observatory (KAO), which is plotted with triangles. The kink seen in the earlier occultation is not evident, indicating a significant change in Pluto's atmospheric structure between 1988 and 2002.

In order to establish exactly what regions of Pluto's atmosphere were probed by our light curves, we determined the UH 2.2-m and Lick half-light times (given in Supplementary Table 2) and then carried out a circular astrometric solution (Fig. 1).

We investigated the extinction properties of Pluto's atmosphere by comparing the minimum normalized stellar fluxes observed in several different spectral bands. Figure 3 shows that the minimum

fluxes increase as a function of wavelength—a behaviour that would not be expected if the minimum flux were determined by refraction (as would be the case for the thermal-gradient model). This trend is just what we would expect, however, from extinction by submicrometre-sized particles¹⁵. In contrast with the 1988 data, the extinction exhibited in the present data does not have an abrupt upper boundary (to which the kink in the light curve has been attributed). One source of submicrometre particles could be photochemical production in Pluto's atmosphere¹⁶. An alternative explanation for the trend seen in Fig. 3 is that we are seeing a small amount of flux from an unresolved, redder star ($K \approx 16.0$ mag) that was

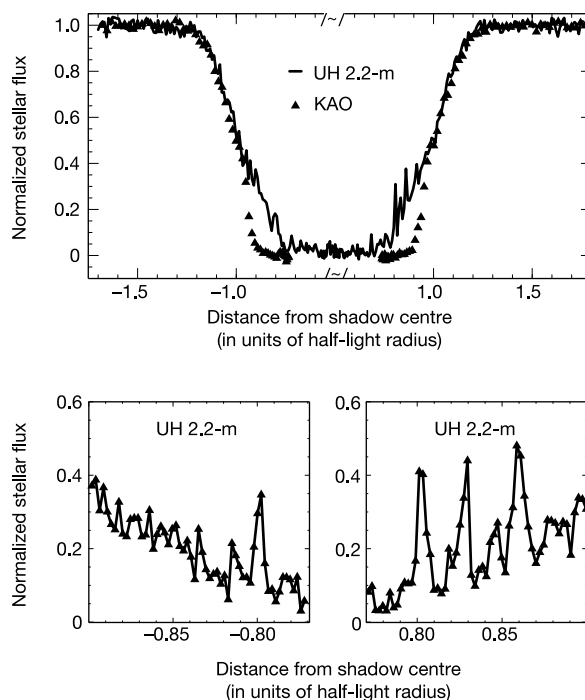


Figure 2 Pluto occultation light curves. Top panel, the UH 2.2-m data from the 21 August 2002 event (line) and the KAO data from the 9 June 1988 event (triangles) have been plotted versus distance from the centre of their occultation shadows. The distance scale has been normalized in units of half-light radius and broken at the centre, making the UH 2.2-m light curve appear as a continuous plot (corresponding to the physical observation of a continuous time series). The KAO data have been averaged over 1.6 s (8 points) and the UH 2.2-m data over 2.0 s (4 points). Since the data have been plotted versus distance from the centre of Pluto's shadow, the central part of the UH 2.2-m light curve exhibits higher frequency variations than would be apparent if the abscissa were time (see, for example, Fig. 1 of the companion paper¹²). The 1988 light curve from the KAO drops sharply to zero just below half light, whereas the 2002 light curve has no

such 'kink'. Since the KAO light curve did not probe as deeply into the shadow as the UH 2.2-m light curve, there is a gap in its centre (although the entire light curve has been plotted). Bottom panels, portions of the UH 2.2-m light curve plotted at their full time resolution of 0.50 s. 'Spikes' (short flashes of starlight) are visible on both the immersion (left) and emersion (right) plots. Such features are barely discernable in the highest time-resolution plots of the KAO data⁵ (not presented here). The spikes are caused by small density variations in Pluto's atmosphere—due to atmospheric waves of some form and/or turbulence—that occur on short spatial scales¹. Note that spikes of similar amplitude appear at 0.8 of a half-light radius from the centre on both the immersion and emersion light curves.

not occulted by Pluto. Such a star could be located as close as ~ 0.03 arcsec (Fig. 1) to P131.1—a possibility that could be checked with high-resolution imaging (a difficult task, even for the Hubble Space Telescope).

Considering the possible pervasive extinction in Pluto's atmosphere, standard techniques for determining the atmospheric structure from an occultation light curve must be used with caution, since these are based on a clear-atmosphere assumption¹⁷ or postulate a specific extinction model². Bearing this in mind, we used only those portions of the light curve corresponding to radii above 1,220 km (to remain above the extinction level in the model used for the 1988 data) to establish the equivalent isothermal temperatures for an N₂ atmosphere. These are 107 ± 4 K and 101 ± 3 K, respectively, for immersion and emersion, and their mean is 104 ± 2 K. This is virtually the same as that obtained in 1988: 104 ± 8 K (the smaller error bar than previously published² reflects a smaller error bar on Pluto's mass¹⁸). Hence the equivalent isothermal temperature of Pluto's atmosphere at the few microbar level has not changed by a detectable amount.

Finally, we inverted the UH 2.2-m light curve (Fig. 2), for which the pressure versus radius profiles are shown in Fig. 4. Here, for comparison, we have also plotted the pressure profiles for the KAO

data from 1988, with the level of the kink indicated for reference. According to the haze model for the 1988 data, all extinction effects should occur below this level (although we cannot determine a maximum altitude for any extinction in the 2002 data). These pressure profiles apply for the clear-atmosphere assumption, but we note that on the scales used in Fig. 4, a break in the pressure profiles at this level (caused by the kink) is not perceptible (in contrast, the shapes of the temperature profiles are highly dependent on the presence of the kink¹⁷). Hence we conclude that the pressure in Pluto's atmosphere within the 1,200–1,280 km radius range has increased by about a factor of two between 1988 and 2002. This corresponds to an expansion of about 40 km for the isobars in the region probed by the occultation.

If the temperature structure between this region and the surface has remained the same, as it apparently has within the region probed by the stellar occultations, then the surface pressure has also increased by a factor of two. Neptune's satellite Triton—a body similar to Pluto in size, density, and atmospheric constituents—experienced a smaller pressure increase^{19,20} during 1989–97. Surface-pressure increases on both of these bodies would be driven by regaining vapour-pressure equilibrium in response to small increases in the temperature (~ 1 K) of nitrogen surface ice. The fact that both have experienced increased pressure leads one to look for a common cause, such as increased solar output—but this has not occurred²¹. Possible reasons for surface-temperature increases include (1) a darkening of the surface ice, so that more sunlight is being absorbed and (2) the effects of thermal inertia and the large obliquity of these bodies. Photometric observations show that Pluto

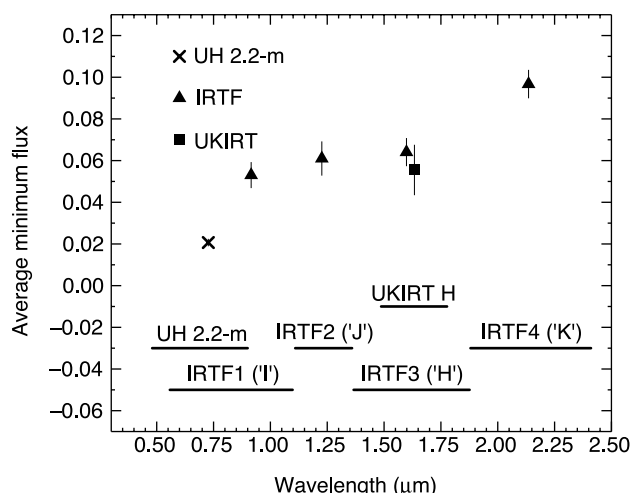


Figure 3 Minimum stellar fluxes versus wavelength. Horizontal lines indicate the pass bands for the UH 2.2-m observations, the H filter used on UKIRT, and the four SpeX 'bands' (crudely approximating the I, J, H and K bands); the points indicating the minimum fluxes are plotted at the effective wavelength for each pass band. Each minimum flux, expressed as a fraction of unocculted stellar flux, and its error bar was derived by averaging the central 216.5 s of the light curve for each data set. The full-scale and zero level of each light curve were calibrated with observations of Pluto-Charon and the star when they were well separated before and after the occultation. The wavelength dependence suggests that the minimum occultation flux is determined by extinction in Pluto's atmosphere. Note that the minimum flux level of the SpeX 'K' band is at about 0.08 of the unocculted stellar flux. By assuming that this flux at K applies to a clear atmosphere, we estimate lower limits on the slant-path optical depths ranging between 0.0 at 'K', through to 0.8 at 'J', and possibly larger than 2 in the visible. Mie¹⁵ light-scattering calculations show that spherical particles with radii near $0.2 \mu\text{m}$ and refractive indices representative of tholins²⁶ have an appropriate wavelength dependence in extinction efficiencies to produce such a curve²⁷. However, modelling of the growth and sedimentation of photochemically produced spherical aerosols²⁸ suggests that realistic haze production rates are not sufficient to produce such large opacities. Whereas single Mie spheres prove inadequate in these models, aggregate particles may provide a solution as they can have extinction efficiencies up to ten times larger than equivalent volume spheres of the same material²⁹. The likelihood of aggregate haze particles in Pluto's atmosphere is strengthened by recent studies^{27,30} showing it is likely that photochemical hazes on Titan are aggregate in nature.

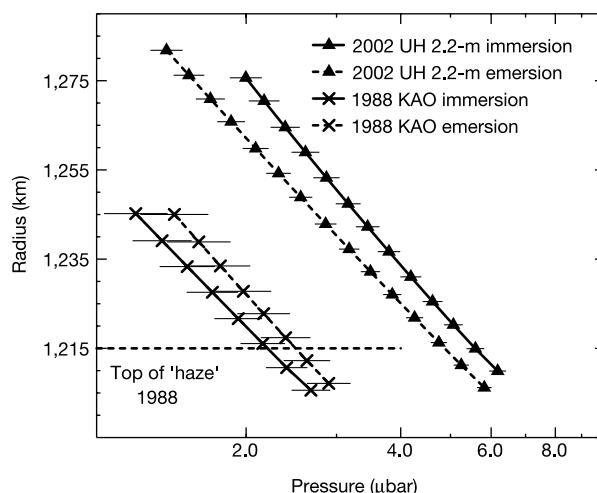


Figure 4 Pressure profiles for 21 August 2002 and 9 June 1988. Pressures obtained from inversions¹⁷ of the two occultation light curves, for immersion and emersion, have been plotted on a log-linear scale versus distance (in km) from the centre of Pluto (the surface radius is poorly constrained¹⁷, with $1,175 \pm 25$ km covering the likely range). The error bars indicate a one standard deviation error, and since these errors are highly correlated¹⁷, the scatter between adjacent points is much less than the error bars. The non-overlapping error bars for the 2002 immersion and emersion pressure profiles may be due to noise or a small discrepancy between the radius scales for immersion and emersion. The large difference between the atmospheric pressures derived from the two occultations indicates that Pluto's atmospheric pressure has increased by a factor of two between 1988 and 2002. A surface-pressure increase by a factor of two on a body with its surface pressure in vapour-pressure equilibrium would imply a temperature increase of the N₂ frost of 1.3 K for the likely range of surface pressures applicable to Pluto¹⁷. If this temperature change in the frost were brought about by changes in the average albedo of Pluto, it would require a decrease in the red albedo, for example, from 0.64 to 0.59. Effects of thermal inertia²² could be at work as well, which would lessen the contribution of additional heat needed from an albedo change.

has been darkening¹¹ since 1954, and certain thermal models predict a large pressure increase at the present epoch²². Although these are plausible explanations, additional data and more modelling will be needed to fully explain our results. □

Received 28 February; accepted 27 May 2003; doi:10.1038/nature01762.

1. Elliot, J. L. & Olkin, C. B. Probing planetary atmospheres with stellar occultations. *Annu. Rev. Earth Planet. Sci.* **24**, 89–123 (1996).
2. Elliot, J. L. & Young, L. A. Analysis of stellar occultation data for planetary atmospheres. I. Model fitting, with application to Pluto. *Astron. J.* **103**, 991–1015 (1992).
3. Brown, G. N. Jr & Ziegler, W. T. Vapor pressure and heats of vaporization and sublimation of liquids and solids of interest in cryogenics below 1-atm pressure. *Adv. Cryog. Eng.* **25**, 662–670 (1980).
4. Hubbard, W. B., Hunten, D. M., Dieters, S. W., Hill, K. M. & Watson, R. D. Occultation evidence for an atmosphere on Pluto. *Nature* **336**, 452–454 (1988).
5. Elliot, J. L. *et al.* Pluto's atmosphere. *Icarus* **77**, 148–170 (1989).
6. Millis, R. L. *et al.* Pluto's radius and atmosphere: Results from the entire 9 June 1988 occultation data set. *Icarus* **105**, 282–297 (1993).
7. Eshleman, V. R. Pluto's atmosphere: Models based on refraction, inversion, and vapor-pressure equilibrium. *Icarus* **80**, 439–443 (1989).
8. Hubbard, W. B., Yelle, R. V. & Lunine, J. I. Nonisothermal Pluto atmosphere models. *Icarus* **84**, 1–11 (1990).
9. McDonald, S. W. & Elliot, J. L. Pluto-Charon stellar occultation candidates: 1996–1999. *Astron. J.* **112**, 788–796 (1996).
10. McDonald, S. W. & Elliot, J. L. Pluto-Charon stellar occultation candidates: 2000–2009. *Astron. J.* **119**, 1999 (2000); Erratum **120**, 1599–1602 (2000).
11. Buie, M. W. *et al.* Changes in Pluto's atmosphere revealed by the P126A occultation. *Bull. Am. Astron. Soc.* **34**, 877 (2002).
12. Sicardy, B. *et al.* Large changes in Pluto's atmosphere as revealed by recent stellar occultations. *Nature* **424**, 168–170 (2003).
13. Rayner, J. T. *et al.* SpeX: A medium-resolution 0.8–5.4 micron spectrograph and imager for the NASA Infrared Telescope Facility. *Publ. Astron. Soc. Pacif.* **115**, 362–382 (2003).
14. Vacca, W. D., Cushing, M. C. & Rayner, J. T. A method for correcting near-infrared spectra for telluric absorption. *Publ. Astron. Soc. Pacif.* **115**, 389–409 (2003).
15. van de Hulst, H. C. *Light Scattering by Small Particles* (Dover, New York, 1981).
16. Summers, M. E., Strobel, D. F. & Gladstone, G. R. in *Pluto and Charon* (eds Stern, S. A. & Tholen, D. J.) 391–434 (Univ. Arizona Press, Tucson, 1997).
17. Elliot, J. L., Person, M. J. & Qu, S. Analysis of stellar occultation data. II. Inversion, with application to Pluto and Triton. *Astron. J.* (in the press).
18. Olkin, C. B., Wasserman, L. H. & Franz, O. G. The mass ratio of Charon to Pluto from Hubble Space Telescope astrometry with Fine Guidance Sensors. *Icarus* (in the press).
19. Elliot, J. L. *et al.* Global warming on Triton. *Nature* **393**, 765–767 (1998).
20. Elliot, J. L. *et al.* The prediction and observation of the 1997 July 18 stellar occultation by Triton: More evidence for distortion and increasing pressure in Triton's atmosphere. *Icarus* **148**, 347–369 (2000).
21. Fröhlich, C. Total solar irradiance variations since 1978. *Adv. Space Res.* **29**, 1409–1416 (2002).
22. Hansen, C. J. & Paige, D. A. Seasonal nitrogen cycles on Pluto. *Icarus* **120**, 247–265 (1996).
23. Buie, M. W. *et al.* CCD camera occultation system. *Bull. Am. Astron. Soc.* **25**, 1115 (1993).
24. Dunham, E. W. in *Proc. Airborne Astronomy Symp. on the Galactic Ecosystem: From Gas to Stars to Dust* (eds Haas, M. R., Davidson, J. A. & Erickson, E. F.) 517–522 (Astronomical Society of the Pacific, San Francisco, 1995).
25. Puxley, P. J. *et al.* Performance of an SBRC 256 × 256 array with ALICE/IRCAM3. *Proc. SPIE* **2198**, 350–358 (1994).
26. Khare, B. N. *et al.* Optical constants of organic tholins produced in a simulated Titanian atmosphere: From soft X-ray to microwave frequencies. *Icarus* **60**, 127–137 (1984).
27. Thomas-Osip, J. E., Elliot, J. L. & Clancy, K. B. Re-examination of the possibility of haze in Pluto's atmosphere based on multi-wavelength observations of the Pluto occultation of P131.1. *Bull. Am. Astron. Soc.* **34**, 1212 (2002).
28. Stansberry, J. A., Lunine, J. I. & Tomasko, M. G. Upper limits on possible photochemical hazes on Pluto. *Geophys. Res. Lett.* **16**, 1221–1224 (1989).
29. Rannou, P., McKay, C. P., Botet, R. & Cabane, M. Semi-empirical model of absorption and scattering by isotropic fractal aggregates of spheres. *Planet. Space Sci.* **47**, 385–396 (1999).
30. Rannou, P. *et al.* Titan's geometric albedo: Role of the fractal structure of the aerosols. *Icarus* **118**, 355–372 (1995).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements Setting up the equipment and recording the occultation data proved to be a challenge, and we appreciate the assistance of G. Koenig, L. Berkgnut, T. Denault, P. Sears, R. Stone, D. Severinsen, A. Pickles and T. Carroll. We thank the staff of the Maui Space Surveillance System for their assistance in taking the Haleakala data, and we thank B. Sicardy *et al.* for communication of results before publication. The United Kingdom Infrared Telescope is operated by the Joint Astronomy Centre on behalf of the UK Particle Physics and Astronomy Research Council. This work was supported, in part, by Research Corporation, the Southwest Research Institute, NSF and NASA. L.C.R. and D.T.H. were supported by Air Force Research Laboratory/Directed Energy.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.L.E. (jle@mit.edu).

Large changes in Pluto's atmosphere as revealed by recent stellar occultations

B. Sicardy^{1,2}, T. Widemann¹, E. Lellouch¹, C. Veillet³, J.-C. Cuillandre³, F. Colas⁴, F. Roques¹, W. Beisker⁵, M. Kretlow⁵, A.-M. Lagrange⁶, E. Gendron¹, F. Lacombe¹, J. Lecacheux¹, C. Birnbaum⁷, A. Fienga⁴, C. Leyrat¹, A. Maury⁸, E. Raynaud¹, S. Renner¹, M. Schultheis⁹, K. Brooks¹⁰, A. Delsanti¹⁰, O. R. Hainaut¹⁰, R. Gilmozzi¹⁰, C. Lidman¹⁰, J. Spyromilio¹⁰, M. Rapaport¹¹, P. Rosenzweig¹², O. Naranjo¹², L. Porras¹², F. Díaz¹², H. Calderón¹², S. Carrillo¹³, A. Carvajal¹³, E. Recalde¹³, L. Gaviria Cavello¹⁴, C. Montalvo¹⁴, D. Barria¹⁵, R. Campos¹⁶, R. Duffard¹⁷ & H. Levato¹⁸

Pluto's tenuous nitrogen atmosphere was first detected by the imprint left on the light curve of a star that was occulted by the planet in 1985 (ref. 1), and studied more extensively during a second occultation event in 1988 (refs 2–6). These events are, however, quite rare and Pluto's atmosphere remains poorly understood, as in particular the planet has not yet been visited by a spacecraft. Here we report data from the first occultations by Pluto since 1988. We find that, during the intervening 14 years, there seems to have been a doubling of the atmospheric pressure, a probable seasonal effect on Pluto.

The stars occulted by Pluto on 20 July 2002 and 21 August 2002 are referred to as P126 and P131.1; respectively, in the candidate list of refs 7 and 8. We organized a campaign for the P126 event, using fixed and portable telescopes in Argentina, Brazil, Chile, Ecuador, Peru and Venezuela, while the P131.1 occultation was observed at the Canada–France–Hawaii Telescope (CFHT) in Hawaii. Independent results obtained from sites in Chile, western continental USA and Hawaii are presented in a companion paper⁹. Owing to weather conditions and astrometric uncertainty before the event, we obtained only one positive detection of the P126 event in northern Chile near Arica, using a portable telescope and a broad-band CCD camera peaking in sensitivity near 0.6 μm (Fig. 1a). In contrast, the P131.1 event was observed with high signal-to-noise ratio in a narrow-band filter at 0.83 μm using the CFHT (Fig. 1b, c).

Our light curves are significantly different from those obtained in 1988. In particular, the Kuiper Airborne Observatory (KAO) data exhibited an abrupt change of slope in their lower part; this change of slope is absent in our data (Fig. 1a, b). This sudden drop was interpreted as being due to either absorbing hazes, or a sharp inversion layer in the lowest 20–50 km above Pluto's surface, connecting the isothermal upper atmosphere at temperature $T \approx 95$ –110 K to the ground temperature at $T \approx 40$ –60 K. Thus, our observations show that large changes in temperature or pressure, or both, occurred in Pluto's lower atmosphere between 1988 and 2002. However, no obvious differences in the shape of the upper part of the light curves are visible, suggesting that the thermal structure of Pluto's upper atmosphere has remained largely unchanged since 1988.

¹Observatoire de Paris, LESIA, 92195 Meudon, France; ²Université Pierre et Marie Curie, 4 Place Jussieu, 75005 Paris, France; ³Canada-France-Hawaii Telescope, Waimea, Hawaii 96743, USA; ⁴Observatoire de Paris, IMCCE, 75014 Paris, France; ⁵International Occultation Timing Association, European Section, 30459 Hanover, Germany; ⁶Observatoire de Grenoble, 38041 Grenoble, France; ⁷Cité des Sciences et de l'Industrie, 75930 Paris, France; ⁸Gene Shoemaker Observatory, Casilla 21, San Pedro de Atacama, Chile; ⁹Institut d'Astrophysique de Paris, 75014 Paris, France; ¹⁰European Southern Observatory, Alonso de Córdova 3107, Casilla 19001, Santiago 19 Chile; ¹¹Observatoire Aquitain des Sciences de l'Univers, 33270 Floirac, France; ¹²Universidad de Los Andes, Facultad de Ciencias, 5101, Mérida, Venezuela; ¹³Cumbaya, 1722 Quito, Ecuador; ¹⁴Asociación Eta Carinae, Lima 1, Perú; ¹⁵Observatório Cerro Armazones, Universidad Católica del Norte, 1280 Antofagasta, Chile; ¹⁶Laboratório Nacional de Astrofísica, 37504-364, Itajubá, Brazil; ¹⁷Observatório Nacional, 20921-400, Rio de Janeiro, Brazil; ¹⁸Complejo Astronómico, El Leoncito, CP J5402DSP, San Juan, Argentina.

Inversions of the light curves in terms of the atmospheric temperature T (Fig. 2a) and pressure p (Fig. 2c) as a function of the distance r to the planet's centre require, for each time step, the distance ρ of the observer to the centre of Pluto's shadow projected on Earth. This distance was retrieved with high accuracy (~ 15 km) for the 1988 event because several occultation chords were available. This, combined with KAO data error bars, provides the 1988 pressure profile shown in Fig. 2c (refs 4, 6). Conversely, with

one chord at hand for the P126 and P131.1 events, the typical uncertainty on ρ is 250 km for both events, as derived from astrometric measurements available for each star and Pluto (see, for example, <http://occult.mit.edu>). From these measurements, the expected closest approaches $\rho_{\min}$ of the Arica and Hawaii stations to the shadow centre were $\rho_{\min} = 975 \pm 250$ km and $\rho_{\min} = 610 \pm 250$ km, respectively.

Only the higher-quality P131.1 light curve has been inverted, and to within the astrometric uncertainties described above, all our inversions show that the temperature or pressure profiles (or both) have drastically changed since 1988. In particular, taking the nominal astrometric solution corresponding to $\rho_{\min} = 610$ km,

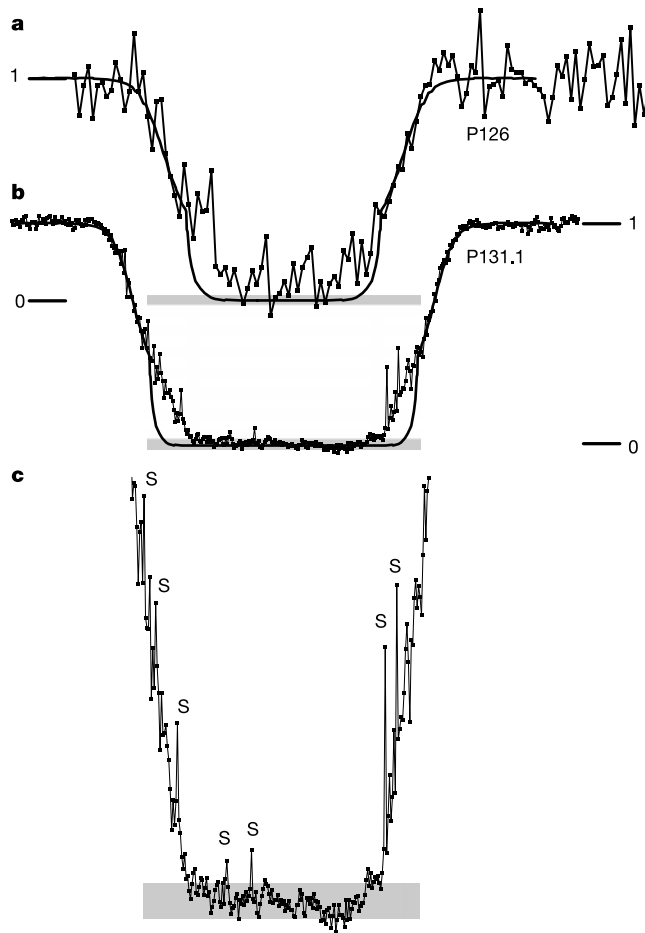


Figure 1 Dimming of the stars P126 (20 July 2002) and P131.1 (21 August 2002) during their occultations by Pluto's atmosphere. **a**, The P126 light curve, binned in 2-s time intervals, was obtained from northern Chile (longitude = $69^{\circ} 45' 51.5''$ W, latitude = $18^{\circ} 26' 53.8''$ S, altitude 2,500 m) with a 30-cm Schmidt–Cassegrain portable telescope and Kodak-401E CCD camera without filter, sensitive between 0.4 and $0.9 \mu\text{m}$, with a peak near $0.6 \mu\text{m}$. The horizontal length of the shaded rectangle represents a duration of 2 min centred at 01:44:03 UT. The total unocculted flux (P126 + Pluto + Charon) is normalized to unity (label "1"). The expected level corresponding to the complete extinction of the star (label "0") is indicated by the position of the shaded rectangle, whose vertical extension represents the uncertainty of photometric calibration. **b**, The P131.1 light curve, obtained at the Canada–France–Hawaii 3.55-m f/8 telescope (CFHT) equipped with a MIT/LL CCD20 in the I band ($0.83 \pm 0.1 \mu\text{m}$) with a exposure time of 1 s and cycle time of 1.583 s. The length of the shaded rectangle now corresponds to a duration of 5 min, centred at 06:50:35 UT. Graphs **a** and **b** are plotted with the same vertical scale from "0" to "1". The two horizontal scales are such that equal horizontal lengths represent equal distances traveled by Pluto in the plane of the sky. Both light curves have absolute timing accuracy of ~ 0.3 s. The solid curve in **a** and **b** is a smoothed version of the data obtained from Kuiper Airborne Observatory (KAO) during the 9 June 1988 Pluto occultation^{3–6}. **c**, A vertically stretched version of **b**, enhancing the spikes ("S") and the steady decrease of signal in the bottom part of the light curve.

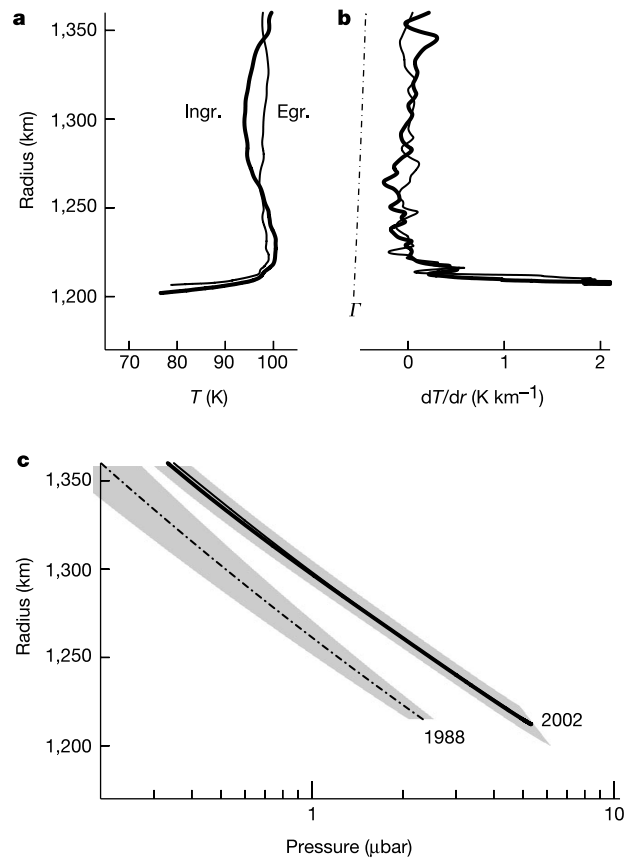


Figure 2 Temperature and pressure profiles of Pluto's atmosphere derived from the inversion of the P131.1 light curve. This inversion¹⁷ assumes a spherically symmetric and transparent atmosphere. It first provides the atmospheric refractivity profile, then the density profile for a given gas composition, and finally the temperature profile, assuming an ideal gas in hydrostatic equilibrium. We assume for Pluto a pure molecular nitrogen⁶ atmosphere, and we take into account the curvature of Pluto's limb as well as the variation of the acceleration of gravity g with radius, assuming a mass of 1.31×10^{22} kg for Pluto¹⁸. **a**, Pluto's atmospheric temperature profiles $T(r)$ at ingress (thick curve) and egress (thin curve). The radius is the distance to the planet centre, for the nominal astrometric solution discussed in the text. **b**, The corresponding temperature gradient profiles dT/dr . The dash-dotted line provides the adiabatic lapse rate profile $\Gamma = -g/c_p$, to the left of which the atmosphere would become convectively unstable, where $c_p = 1.04 \times 10^3 \text{ J K}^{-1} \text{ kg}^{-1}$ is the specific heat at constant pressure for N_2 . **c**, The corresponding pressure profiles. The dash-dotted line is the pressure profile derived from the KAO lightcurve of 9 June 1988 (ref. 6), with the associated error domain⁶, shown as the shaded region. The ingress and egress profiles derived from the CFHT light curve of 2002 are the nearly coincident thick and thin solid lines, respectively. The error domain for the 2002 profiles only takes into account the photometric uncertainties in the levels "0" and "1" shown in Fig. 1b. The effects of the astrometric uncertainty are discussed in the text.

we obtain essentially the same temperature profile $T(r)$ as in 1988, including a strong inversion layer below $r = 1,215$ km, but with a systematic and significant increase of pressure at all radii (Fig. 2c). This pressure increase reaches a factor of 2.1 at $r = 1,215$ km, where $p = 5.0 \pm 0.6 \mu\text{bar}$ in 2002, while $p = 2.33 \pm 0.24 \mu\text{bar}$ in 1988 (ref. 6).

Figure 1c shows that the bottom of the P131.1 light curve is generally flat for about 200 s, with a small but significant decrease of signal over this time interval. Taking again $\rho_{\min} = 610$ km, we find that during this 200-s interval the image of P131.1 scanned Pluto's limb at a constant radius of 1,205 km, with latitudes ranging from 61°S (for the onset of occultation, or 'ingress', also known as 'immersion') to 8°S (for the termination of occultation, 'egress' or 'emersion'). This steady decrease of signal could conceivably be due to the presence of morning hazes above the egress point. Perhaps more probably, it indicates an egress temperature about 10 K smaller than the ingress temperature at $r = 1,205$ km. The P131.1 star ingress probed the south pole region now constantly in sunlight, while the egress probed the morning limb of the equatorial region, which emerged after 3.2 days in darkness. (Note that we adopt the International Astronomical Union convention, in which Pluto's south pole is currently tilted toward Earth). Taking a thermal inertia of $15 \text{ J m}^{-2} \text{ s}^{-1/2} \text{ K}^{-1}$, derived from Pluto's thermal light curve¹⁰, and using a subsolar latitude of 29.7°S at Pluto in August 2002, we estimate that the dawn equatorial ground temperature is $\sim 25\%$ lower, that is, 12–15 K cooler, than the polar temperature for a uniform surface albedo. However, according to a recent surface model¹¹, Pluto's south pole is covered by methane-rich areas while the equator is predominantly covered by darker tholins. For this case, we estimate a ground polar temperature of ~ 53 K and a dawn equatorial temperature of 45 K. These numbers compare well with the difference between the ingress and egress temperatures near $r = 1,205$ km, suggesting that the lower 10–40 km (0.25–1 scale height) of the atmosphere are affected by surface boundary layer effects.

Conversely, our observations do not reveal a significant temperature difference between egress and ingress (and also with 1988) in the isothermal region above $r = 1,220$ km. This can be understood in the methane-thermostat model of ref. 12, which demonstrates that the upper-atmosphere temperature is robustly controlled at ~ 100 K by the radiative properties of atmospheric CH_4 , almost independently of its abundance. Accounting for the diurnally averaged insolation at the south pole and the equator, we find that the upper-atmosphere temperatures at ingress and egress should only differ by 3 K at radiative equilibrium, well within the uncertainties of our measurements.

The P131.1 light curve exhibits spikes (Fig. 1c) which correspond to local temperature gradients of up to $dT/dr \approx \pm 0.05 \text{ K km}^{-1}$ over vertical scales of 5–15 km in the retrieved temperature profiles (Fig. 2b), yielding temperature fluctuations of $\Delta T \approx \pm 0.5\text{--}0.8$ K over these scales. The spikes are more conspicuous than in 1988 (refs 4, 6), and reveal a dynamical activity in Pluto's atmosphere. Its origin could be gravity waves launched by convection in a putative lower troposphere, and/or turbulence triggered by shears driven by strong winds between Pluto's lit and dark hemispheres. The observed values of $|dT/dr|$ remain well below the adiabatic temperature lapse rate Γ (Fig. 2b), indicating that either convective instability is not reached, or that stronger local gradients exist but are smoothed out over the ~ 500 km that a stellar ray travels in the atmosphere before being significantly refracted towards Earth.

We have tested the robustness of our results to changes in the P131.1 astrometric solution. We find that varying $\rho_{\min}$, everything else being equal, mainly displaces vertically the temperature profile $T(r)$, while multiplying the pressure at all levels by a roughly constant factor with respect to the nominal solution with $\rho_{\min} = 610$ km. In particular, a solution that leaves the pressure

unchanged at all levels (to within 10%) since 1988 can be found with $\rho_{\min} = 550$ km. However, the whole temperature profile $T(r)$, including the inversion layer, would then be shifted downward by about 30 km. Such a large variation of Pluto's thermal structure in the lower atmosphere since 1988 seems unlikely. The inversion layer is best physically explained through the absorption by atmospheric methane, and its existence does not depend sensitively on the CH_4 abundance (although the precise value of the temperature gradient does¹³). Moving the inversion layer downward by 30 km from its $r = 1,200\text{--}1,220$ km nominal level would imply a dramatic shrinking of the 40-km-deep troposphere advocated in ref. 14 on the basis of the 1988 occultation. Such a troposphere would presumably be maintained by convection, perhaps driven by the large temperature gradients across Pluto's surface¹⁰. It should be fairly stable over a 14-year timescale, and it is supported by the presence of spikes in the CFHT light curve, as discussed above.

In contrast, a twofold increase in pressure in the 14 years since 1988 is not at odds with expectations of nitrogen cycle models^{15,16}, despite Pluto's increasing heliocentric distance. The best-fitting models¹⁵, in which N_2 frost distribution and associated atmospheric pressure are calculated as a function of season, do predict an increase of pressure by a factor of about 2 between 1990 and 2000; this increase follows equinox on Pluto in 1987, leading to the current sublimation of the south polar cap as it goes into sunlight. Thus, although the seasonal models are only partly successful in reproducing the observed albedo patterns—particularly the currently large north polar cap and the relative darkness of the south polar regions, they must adequately capture the physics of hemispheric volatile exchange. $\square$

Received 4 March; accepted 27 May 2003; doi:10.1038/nature01766.

1. Brosch, N. The 1985 stellar occultation by Pluto. *Mon. Not. R. Astron. Soc.* **276**, 571–578 (1995).
2. Hubbard, W. B., Hunten, D. M., Dieters, S. W., Hill, K. M. & Watson, R. D. Occultation evidence for an atmosphere on Pluto. *Nature* **336**, 452–454 (1988).
3. Elliot, J. L. *et al.* Pluto's atmosphere. *Icarus* **77**, 148–170 (1989).
4. Elliot, J. L. & Young, L. A. Analysis of stellar occultation data for planetary atmospheres. I. Model fitting, with application to Pluto. *Astron. J.* **103**, 991–1051 (1992).
5. Millis, R. L. *et al.* Pluto's radius and atmosphere—Results from the entire 9 June 1998 occultation data set. *Icarus* **105**, 282–297 (1993).
6. Yelle, R. V. & Elliot, J. L. in *Pluto and Charon* (eds Stern, S. A. & Tholen, D. J.) 347–390 (Univ. Arizona Press, Tucson, 1997).
7. McDonald, S. W. & Elliot, J. L. Pluto-Charon stellar occultation candidates: 2000–2009. *Astron. J.* **119**, 1999–2007 (2000).
8. McDonald, S. W. & Elliot, J. L. Erratum: "Pluto-Charon stellar occultation candidates: 2000–2009". *Astron. J.* **120**, 1599–1602 (2000).
9. Elliot, J. L. *et al.* The recent expansion of Pluto's atmosphere. *Nature* **424**, 165–168 (2003).
10. Lellouch, E. *et al.* Pluto's non-isothermal surface. *Icarus* **147**, 220–250 (2000).
11. Grundy, W. M. & Buie, M. W. Distribution and evolution of CH_4 , N_2 and CO ices on Pluto's surface: 1995 to 1998. *Icarus* **153**, 248–263 (2001).
12. Yelle, R. V. & Lunine, J. I. Evidence for a molecule heavier than methane in the atmosphere of Pluto. *Nature* **339**, 288–290 (1989).
13. Strobel, D. F., Zhu, X., Summers, M. E. & Stevens, M. H. On the vertical thermal structure of Pluto's atmosphere. *Icarus* **120**, 266–289 (1996).
14. Stanberry, J. A., Lunine, J. I., Hubbard, W. B., Yelle, R. V. & Hunten, D. M. Mirages and the nature of Pluto's atmosphere. *Icarus* **111**, 503–513 (1994).
15. Hansen, C. J. & Paige, D. A. Seasonal nitrogen cycles on Pluto. *Icarus* **120**, 247–265 (1996).
16. Spencer, J. R., *et al.* in *Pluto and Charon* (eds Stern, S. A. & Tholen, D. J.) 435–475 (Univ. Arizona Press, Tucson, 1997).
17. Sicardy, B. *et al.* The structure of Titan's atmosphere from the 28 Sgr occultation. *Icarus* **142**, 357–390 (1999).
18. Null, G. W. & Owen, W. M. Jr Charon/Pluto mass ratio obtained with *HST* CCD observations in 1991 and 1993. *Astron. J.* **111**, 1368–1381 (1996).

Acknowledgements We thank the Conseil Scientifique of Paris Observatory and the Programme National de Planétologie for supporting part of the observations of the P126 event in South America. We are grateful to C. Angeli, A. Bruch and D. Lazzaro, who made possible the observational campaign in Brazil, and to W. B. Hubbard for constructive comments.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.S. (bruno.sicardy@obspm.fr).

Miniaturized gas ionization sensors using carbon nanotubes

Ashish Modi*, Nikhil Koratkar*, Eric Lass†, Bingqing Wei† & Pulickel M. Ajayan†

* Department of Mechanical, Aerospace and Nuclear Engineering, and
† Department of Materials Science and Engineering, Rensselaer Polytechnic Institute, Troy, New York 12180, USA

Gas sensors operate by a variety of fundamentally different mechanisms^{1–14}. Ionization sensors^{13–14} work by fingerprinting the ionization characteristics of distinct gases, but they are limited by their huge, bulky architecture, high power consumption and risky high-voltage operation. Here we report the fabrication and successful testing of ionization microsensors featuring the electrical breakdown of a range of gases and gas mixtures at carbon nanotube tips. The sharp tips of nanotubes generate very high electric fields at relatively low voltages, lowering breakdown voltages several-fold in comparison to traditional electrodes, and thereby enabling compact, battery-powered and safe operation of such sensors. The sensors show good sensitivity and selectivity, and are unaffected by extraneous factors such as temperature, humidity, and gas flow. As such, the devices offer several practical advantages over previously reported nanotube sensor systems^{15–17}. The simple, low-cost, sensors described here could be deployed for a variety of applications, such as environmental monitoring, sensing in chemical processing plants, and gas detection for counter-terrorism.

Various groups have explored the potential of carbon nanotubes for gas sensing, based on the electrical conductance changes of semiconducting nanotubes on exposure to gases^{15–17}. Although this method has high sensitivity, it is limited by several factors, such as the inability to identify gases with low adsorption energies, poor diffusion kinetics or poor charge transfer with nanotubes. It is also challenging to use this technique to distinguish between gases or gas mixtures; gases in different concentrations could produce the same net change in conductance as produced by a single pure gas. Nanotube conductance is also very sensitive to changes in environmental conditions (moisture, temperature, gas-flow velocity), and chemisorption could cause irreversible changes in nanotube conductivity.

Figure 1 shows details of our sensor—a diagram of the component parts (Fig. 1a), and the configuration of the electrodes (Fig. 1b). Controlled d.c. voltage is applied between the anode (vertically aligned multiwalled nanotube—MWNT—film) and the cathode (Al sheet), which are separated by a glass insulator. The MWNT film (Fig. 1c) used for this test was grown by chemical vapour deposition^{18–20} (CVD) on an SiO₂ substrate, as reported earlier. Nanotubes in the film are ~25–30 nm in diameter, ~30 µm long, with ~50 nm separation between nanotubes²⁰.

Individual MWNTs within the film, owing to their nanometre-scale tip radius ($R \approx 15$ nm), create very high nonlinear electric fields near the tips^{21–24}. This hastens the breakdown process due to formation of a ‘corona’ or conducting filament of highly ionized gas that surrounds the MWNT tips. This corona promotes the formation of a powerful electron avalanche or plasma streamer that bridges the gap between the electrodes, and allows a self-sustaining interelectrode discharge to be created at relatively low voltages. The technique is very powerful for several reasons. (1) The precise breakdown voltage provides the ‘fingerprint’ for the gas to be identified; it is well established that at constant temperature and pressure every gas has a unique^{25,26} breakdown electric field. (2) By monitoring the self-sustaining discharge current, the gas concentration can be determined. (3) Because this technique does not

involve adsorption/desorption of gases, the sensor displays a fast response and is not limited by considerations of reversibility.

The device was first tested in air (Fig. 2a) with anode–cathode separation of 150 µm. Continuous current discharge of 460 µA was generated at 346 V. Tests were repeated with metal electrodes (no nanotubes), while still maintaining the electrode separation of 150 µm. For this case, breakdown voltage of air occurred at 960 V with current discharge of 69 µA. This shows that by the use of MWNTs as the anode, the breakdown voltage of air can be reduced nearly 65%. The discharge current is also increased from 69 µA to 460 µA (~6-fold increase), leading to high sensitivity (discharge current at breakdown indicates concentration of the detected species, therefore a high discharge current enables detection of dilute concentrations of the unknown gas). We believe that the observed increase in discharge current is related to the high density of MWNTs that constitute the electrode surface (Fig. 1c). The billions of aligned nanotubes covering the substrate produce a consistent nanometre-scale surface topology unobtainable for conventional planar electrodes or micromachined electron emitters. A significant number of these tubes are expected to participate in ionization, leading to a more extensive conduction path and consequently higher discharge current. The devices were robust, with no degradation observed after hours of operation.

This nanotube-anode ionization device was used to detect the identity of several gas species, such as helium, argon, nitrogen, oxygen, carbon dioxide, ammonia and air. The nanotube device was placed in an environmental chamber with electrical feed-through, and air was pumped out of the chamber to establish a high vacuum (10^{-4} torr). The gas to be identified was then released in a controlled fashion. Breakdown data were recorded over a wide range of gas concentrations (10^{-7} to 10^{-1} mol l⁻¹). Figure 2b shows the breakdown voltages of several gases at room temperature (300 °K) and at a chamber pressure of 760 torr, that is, a gas concentration of 4×10^{-2} mol l⁻¹. For all tests shown, the anode–cathode separation was maintained at 150 µm. Note that each gas exhibits a distinct breakdown behaviour; helium displays the lowest (164 V) and ammonia shows the highest (430 V) breakdown voltage.

To study the effect of gas concentration, tests were conducted at

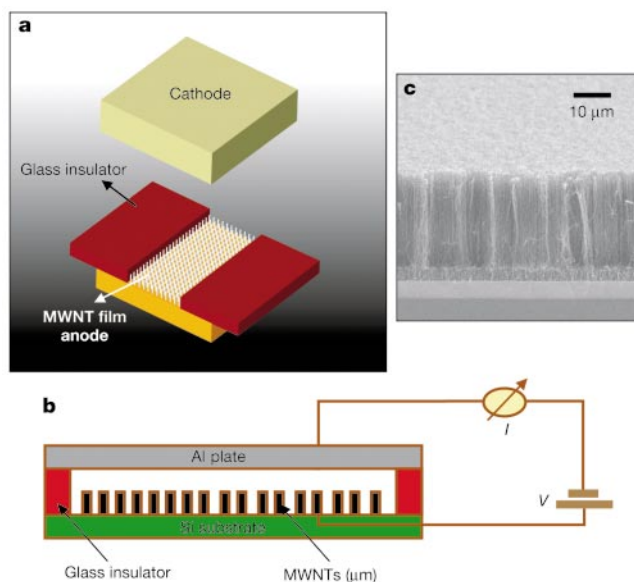


Figure 1 The nanotube sensor device. **a**, Exploded view of sensor showing MWNT film as the anode, 180-µm-thick glass insulator plates, and Al plate as cathode. **b**, Diagram of actual test set-up. **c**, SEM micrograph of a CVD-grown, vertically aligned MWNT film used as the anode.

reduced pressures (Fig. 3a, b). Figure 3a shows the effect of concentration on the breakdown voltages of air, argon, helium and ammonia. Note that the breakdown voltage does not vary significantly with gas pressure (the pressure is proportional to concentration for a fixed chamber volume). This is because breakdown behaviour in this case is dominated by the highly nonlinear electric field near the nanotube tips, resulting in a pre-breakdown plasma that helps to bridge the electrode gap and reduces the sensitivity of breakdown voltage to gas pressure. But at very low gas concentrations (below $10^{-6} \text{ mol l}^{-1}$), the breakdown voltage did increase as predicted by Paschen's law for uniform electric field^{25,26}, indicating a certain concentration threshold needed for the discharge to be self-sustaining. From Figs 2b and 3a we conclude that, for a fixed interelectrode spacing of the device, the breakdown voltage of each gas is unique and depends mainly on the electric field, being only weakly affected by concentration (this is valid over a wide range; Fig. 3a). Therefore by monitoring the breakdown voltage of the gas, its identity can be established. Note that species such as NH_3 and He that have large separation in breakdown voltage (Fig. 2b) are easier to distinguish than, say, NH_3 and O_2 .

Figure 3b shows the self-sustaining current discharge at break-

down for argon, nitrogen, oxygen, ammonia and air. The discharge current varies logarithmically with concentration. This trend is valid over a wide range of gas concentrations, ranging from 10^{-7} to $10^{-1} \text{ mol l}^{-1}$. This shows that the self-sustaining discharge current generated at breakdown is a characteristic property of the number of gas molecules per unit volume that are available for conduction. For example, Fig. 3b indicates that for N_2 a current discharge of $328 \mu\text{A}$ corresponds to a concentration of $3.22 \times 10^{-5} \text{ mol l}^{-1}$. The discharge current increases logarithmically to about $420 \mu\text{A}$ as the N_2 concentration is increased to $4 \times 10^{-2} \text{ mol l}^{-1}$. Therefore the discharge current provides a convenient means to quantify the concentration of the species being detected. Figure 3b also shows that no hysteresis is observed in the sensor response, even for species such as NH_3 and O_2 that are known to interact strongly¹⁵⁻¹⁷ with nanotube surfaces at room temperature.

The breakdown process described earlier propagates through the formation of a positive or negative corona, depending on whether the MWNT film is configured as the anode or the cathode. For the experiments with the MWNT film as anode, the breakdown propagation mechanism is via the formation of a positive corona. We also conducted experiments in air with the MWNT film as

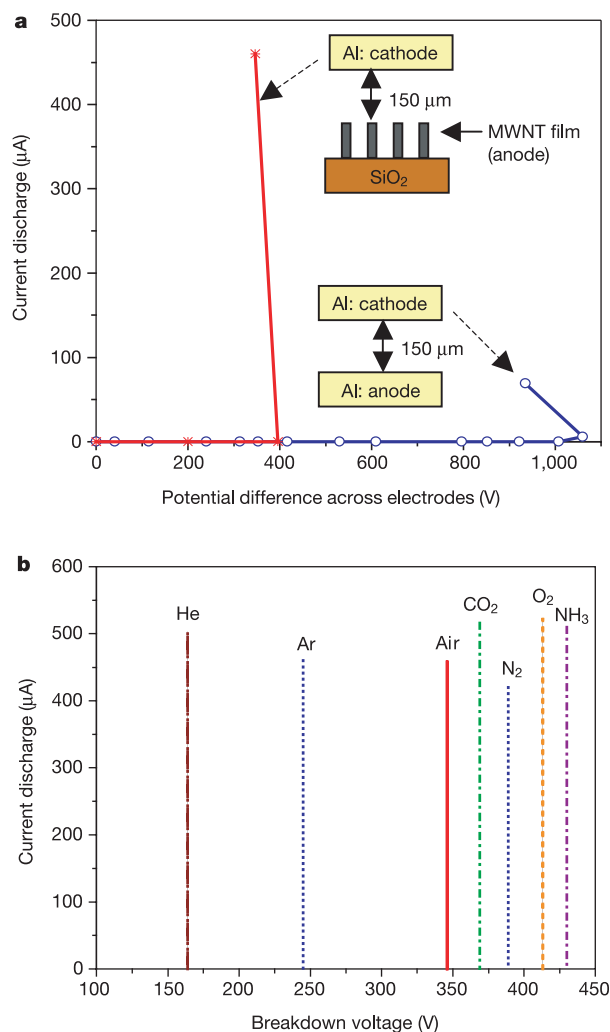


Figure 2 Current–voltage (I – V) curves for electrical breakdown. **a**, I – V curve (blue) with Al plate as the anode, separated from cathode by $150 \mu\text{m}$, showing breakdown voltage of air at 960 V with discharge current of $69 \mu\text{A}$; with MWNT film as anode (red) showing breakdown voltage of air at 346 V with discharge current of $460 \mu\text{A}$. **b**, I – V curves for NH_3 , CO_2 , N_2 , O_2 , He , Ar and air, showing distinct breakdown voltages; ammonia displays the highest breakdown voltage, and helium the lowest.

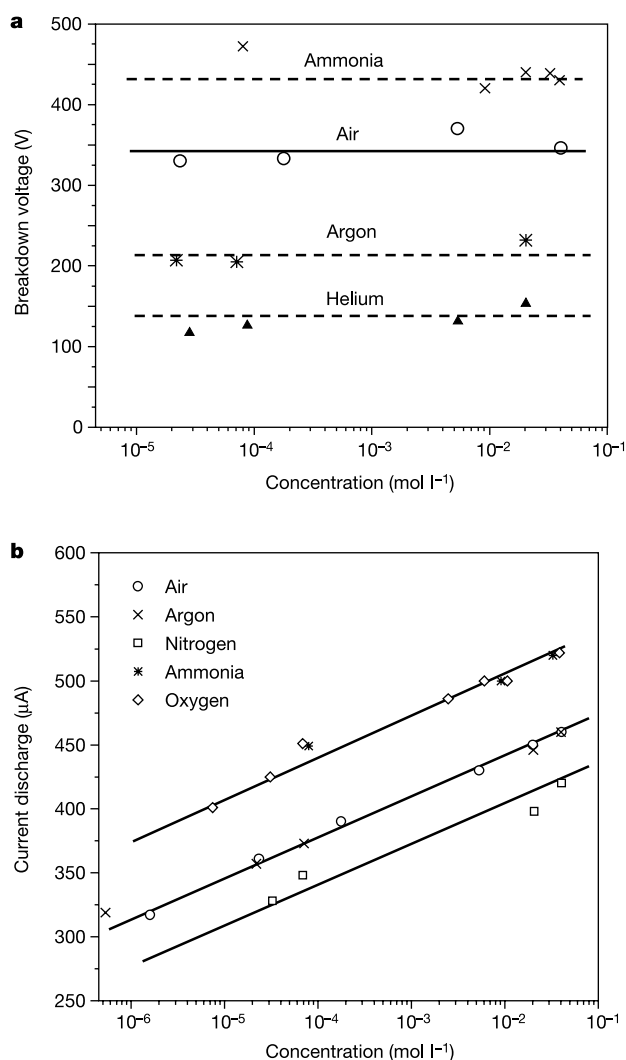


Figure 3 Effect of gas concentration on electrical breakdown. **a**, Breakdown voltage as a function of concentration; breakdown voltages vary only slightly with gas concentration. **b**, Discharge current at breakdown as a function of gas concentration. The discharge current varies logarithmically with concentration.

cathode (negative corona). The tests indicate that the voltage at which breakdown is initiated is similar for both positive (350 V) and negative (330 V) corona; but in negative corona, the breakdown voltage decreases to a lower value (~ 280 V) once breakdown is established. We believe that this is because once ionization is initiated in a negative corona, free electrons are repelled away from nanotube tips into the surrounding gas, triggering secondary ionizations that enable a self-propagating electron avalanche (a conducting path) to form at a lower electric field.

The breakdown voltage was found to become lower as the interelectrode spacing was reduced. This is expected, as reducing the electrode separation increases the electric field in the gap. Figure 4a shows breakdown voltage as a function of electrode separation. Two cases are shown, one when both electrodes were made out of Al plates and the other with the MWNT film as cathode. For the Al cathode, breakdown voltages came down from 1,050 V at 150 μm separation to 354 V at 28 μm separation; for the MWNT-film cathode, breakdown voltages went down from 280 V (at 150 μm separation) to 130 V (at 25 μm separation). Such voltages can be easily obtained by connecting six 22.5 V carbon–zinc batteries in series, suggesting that portable nanotube sensors

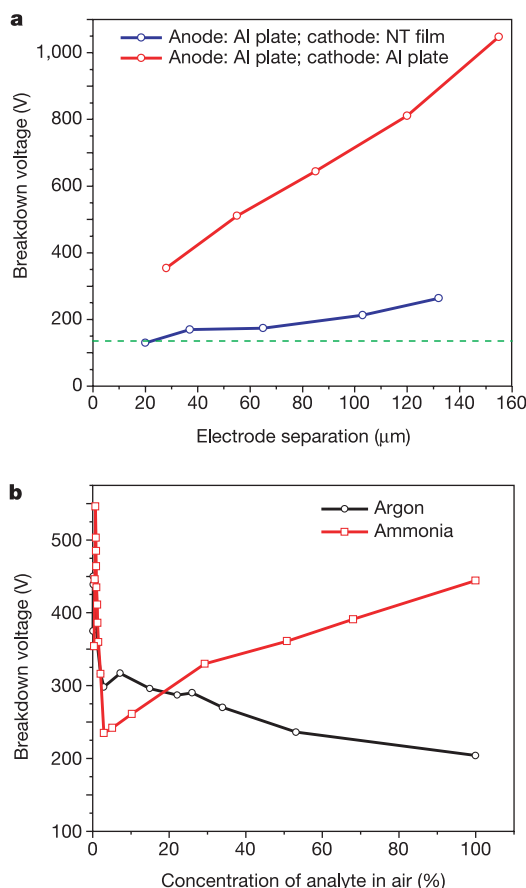


Figure 4 Effect of electrode separation, and detection of gases in mixtures. **a**, Breakdown voltage of air versus interelectrode separation. Breakdown voltages reduced from 280 V at 150 μm separation, to 130 V at 25 μm separation, for an MWNT cathode, and reduced from 1,050 V at 150 μm separation, to 354 V at 28 μm separation, for an Al cathode. Green line indicates the voltage obtained by connecting six, 22.5 V carbon–zinc batteries (Eveready AAA) in series. **b**, Breakdown of Ar and NH_3 in a mixture with air (MWNT film is configured as the anode, and interelectrode gap is 150 μm). Breakdown voltage increases with decreasing Ar concentration in the mixture. Breakdown of Ar ceases at about 1% concentration, and the breakdown voltage ramps up to that of pure air. For NH_3 , the breakdown voltage decreases with decreasing NH_3 concentration until about 1%, beyond which the breakdown behaviour is identical to that of pure air.

based on the above idea could be made and utilized.

Ionization sensors such as photo-ionization detectors (PID), flame-ionization detectors (FID) or electron-capture detectors (ECD) are not suitable for direct application to gas mixtures. These detectors work in conjunction with a gas-chromatography set-up that separates the mixture into distinct bands that can then be qualitatively and quantitatively analysed. The nanotube ionization sensor can be used to monitor gas mixtures without the direct use of a chromatography arrangement. Figure 4b shows the results for an Ar–air mixture with several different relative concentrations of the component gases. As expected, for over 50% Ar in the mixture, the breakdown voltage is nearly the same as that of pure Ar. As the relative concentration of Ar in the mixture is reduced, the breakdown voltage increases from about 250 V (for 50% Ar) to about 300 V (for 1% Ar). This is because air has a higher breakdown voltage than Ar, so the presence of air molecules tends to impede the breakdown of Ar. Below 1% concentration, the breakdown of Ar ceases and the breakdown voltage rises sharply to the value for pure air (~ 350 V). Similar results (detection limits $\sim 1\%$) were also obtained for detection of NH_3 in a mixture with air (Fig. 4b). These tests indicate that the nanotube ionization microsensor (with proper calibration) shows promise for room-temperature detection of gases at the percentage level in mixtures with air with fast response (application of the breakdown electric field results in a stable discharge within ~ 20 μs ; ref. 26). In contrast, the metal oxide sensors² used in industry typically operate at 300–500 $^\circ\text{C}$ for detection of 1% NH_3 with a response time of ~ 1 minute. Conducting polymer sensors⁴ show detection limits of 1% for NH_3 at room temperature, but require ~ 10 minutes for sensing.

The nanotube ionization sensor can also be used as a detector in gas chromatographs, where sufficiently large analyte concentrations are at hand. In fact, the nanotube ionization detector could offer several advantages over the FID, PID or ECD detectors that are routinely used in chromatography sensors. FID has poor selectivity and requires bulky and hazardous hydrogen storage tanks during operation, PID with a better selectivity is limited to a small range of analytes, and ECD detectors are hazardous because they contain radioactive electron emitters. In contrast, the nanotube sensor is compact, safe to use and requires low power to operate. Since every gas has a characteristic breakdown electric field, a gas chromatograph with a nanotube ionization detector could potentially be applied to a broad range of analytes with good selectivity (Figs 2b, 3a).

To investigate the sensitivity range of a nanotube-chromatography sensor, we conducted a simulated gas-chromatography test with He chosen as the mobile phase. The results (Supplementary Information) indicate that, with appropriate design of the chromatography arrangement (including choice of mobile phase, stationary phase and process parameters), detection of analytes in the low p.p.m. range (~ 25 p.p.m.) appears feasible with this technique. Compact, low-power nanotube detectors coupled to miniature separation columns could potentially lead to truly field-portable gas chromatographs that could be used during emergency response and counter-terrorism situations that require definitive identification of contaminants in near real-time. $\square$

Received 27 January; accepted 28 May 2003; doi:10.1038/nature01777.

- Watson, J. & Ihokura, K. (eds) Special issue on gas sensing materials. *Mater. Res. Soc. Bull.* **24** (1999).
- Shimizu, Y. & Egashira, M. Basic aspects and challenges of semiconductor gas sensors. *Mater. Res. Soc. Bull.* **24**, 18–24 (1999).
- Takao, Y., Miyazaki, K., Shimizu, Y. & Egashira, M. High ammonia sensitive semiconductor gas sensors with double-layer structure and interface electrodes. *J. Electrochem. Soc.* **141**, 1028–1034 (1994).
- Miasik, J., Hooper, A. & Tofield, B. Conducting polymer gas sensor. *J. Chem. Soc. Faraday Trans.* **182**, 1117–1126 (1986).
- Capone, S., Mongelli, S., Rella, R., Siciliano, P. & Valli, L. Gas sensitivity measurements on NO_2 sensors based on copper (II) tetrakis (n-butylaminocarbonyl) phthalocyanine LB films. *Langmuir* **15**, 1748–1753 (1999).
- Endres, H., Hartinger, R., Schwaiger, M., Gmelch, G. & Roth, M. A capacitive CO_2 sensor system with suppression of the humidity interference. *Sens. Actuators B* **57**, 83–87 (1999).

7. Lee, M. & Meyer, J. A new process for fabricating CO₂ sensing layers based on BaTiO₃ and additives. *Sens. Actuators B* **68**, 293–299 (2000).
8. Matsubara, S. *et al.* A practical capacitive type CO₂ sensor using CeO₂/BaCO₃/CuO ceramics. *Sens. Actuators B* **65**, 128–132 (2000).
9. Currie, J., Essalik, A. & Marusic, J. Micromachined thin film solid state electrochemical CO₂, NO₂ and SO₂ gas sensors. *Sens. Actuators B* **59**, 235–241 (1999).
10. Longergan, M. *et al.* Array-based vapor sensing using chemically sensitive, carbon black-polymer resistors. *Chem. Mater.* **8**, 2298–2312 (1996).
11. Dresselhaus, M., Dresselhaus, G. & Avouris, P. Carbon nanotubes — synthesis, structure, properties and applications. *Top. Appl. Phys.* **80**, 391–425 (2001).
12. Baughman, R. H., Zakhidov, A. A. & de Heer, W. A. Carbon nanotubes—the route toward applications. *Science* **297**, 787–792 (2002).
13. The Environmental Response Team. *Standard Operating Procedure No. 2114* (Environmental Protection Agency, Washington DC) (<http://www.ertresponse.com/sops/2114.pdf>) (1994).
14. Flame ionization detector, product data sheet (SRI Instruments, Torrance, CA) (<http://www.srigc.com/FID.pdf>) (1998).
15. Collins, P., Bradley, K., Ishigami, M. & Zettl, A. Extreme oxygen sensitivity of electronic properties of carbon nanotubes. *Science* **287**, 1801–1804 (2000).
16. Kong, J. *et al.* Nanotube molecular wires as chemical sensors. *Science* **287**, 622–625 (2000).
17. Peng, S., *et al.* in *3rd Int. Workshop on Structural Health Monitoring* (ed. Chang, F. K.) 1142–1148 (CRC Press, 2001).
18. Zhang, Z., Wei, B. Q., Ramanath, G. & Ajayan, P. M. Substrate-site selective growth of aligned carbon nanotubes. *Appl. Phys. Lett.* **77**, 3764–3766 (2000).
19. Wei, B. Q. *et al.* Organized assembly of carbon nanotubes. *Nature* **416**, 495–496 (2002).
20. Drotar, J. *et al.* Reflection high-energy electron diffraction from carbon nanotubes. *Phys. Rev. B* **64**, 125417 (2001).
21. de Heer, W. A., Chatelain, A. & Ugarte, D. A carbon nanotube field-emission electron source. *Science* **270**, 1179–1180 (1995).
22. de Jonge, N., Lamy, Y., Schoots, K. & Oosterkamp, T. H. High brightness electron beam from a multi-walled carbon nanotube. *Nature* **420**, 393–395 (2002).
23. Forbes, R. G., Edgcombe, C. J. & Valdre, U. Some comments on models for field enhancement. *Ultramicroscopy* **95**, 57–65 (2003).
24. Edgcombe, C. J. & Valdre, U. Microscopy and computational modeling to elucidate the enhancement factor for field electron emitters. *J. Microsc.* **203**, 188–194 (2001).
25. Meek, J. & Craggs, J. *Electrical Breakdown of Gases* (Wiley and Sons, New York, 1978).
26. Abdel-Salam, M., *et al.* *High Voltage Engineering—Theory and Practice* (Dekker, New York, 2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This work was supported by NSF (NSEC, NER awards), US Army Research Office, Philip Morris USA, and RPI.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to N.K. (koratn@rpi.edu) and P.M.A. (ajayan@rpi.edu).

Unidirectional rotation in a mechanically interlocked molecular rotor

David A. Leigh*, Jenny K. Y. Wong*, François Dehez† & Francesco Zerbetto†

* School of Chemistry, University of Edinburgh, The King's Buildings, West Mains Road, Edinburgh EH9 3JJ, UK

† Dipartimento di Chimica 'G. Ciamician', Università degli Studi di Bologna, via F. Selmi 2, 40126 Bologna, Italy

Molecular motor proteins are ubiquitous in nature¹ and have inspired attempts to create artificial machines² that mimic their ability to produce controlled motion on the molecular level. A recent example of an artificial molecular rotor is a molecule undergoing a unidirectional 120° intramolecular rotation around a single bond^{3,4}; another is a molecule capable of repetitive unimolecular rotation driven by multiple and successive isomerization of its central double bond^{5–8}. Here we show that sequential and unidirectional rotation can also be induced in mechanically interlocked assemblies comprised of one or two small rings moving around one larger ring. The small rings in

these [2]- and [3]catenanes⁹ move in discrete steps between different binding sites located on the larger ring, with the movement driven by light, heat or chemical stimuli that change the relative affinity of the small rings for the different binding sites^{10–12}. We find that the small ring in the [2]catenane moves with high positional integrity but without control over its direction of motion, while the two rings in the [3]catenane mutually block each other's movement to ensure an overall stimuli-induced unidirectional motion around the larger ring.

A [2]catenane in which one ring moves sequentially between three different binding sites ('stations') on the other ring is shown schematically in Fig. 1. A series of chemical reactions alters the thermodynamics of binding of two of the binding sites (switching them 'on' or 'off' in response to external stimuli) causing the small ring to move around the larger ring in successive steps in response to the changing global energy minimum. A hydrogen-bonded [2]catenane, **1**, based on this design was prepared together with the corresponding macrocycle, **2**, and [3]catenane (two of the small rings linked onto the bigger ring), **3**, as outlined in Fig. 2 and the Supplementary Information. The stations where the small rings non-covalently bind to the large ring were predicted to have widely differing macrocycle binding affinities (Fig. 3A): **A**, a secondary (2°) amide fumaramide group, should bind strongly because the *E*-olefin holds the amide carbonyls in a close-to-ideal geometry for intercomponent hydrogen bonding¹⁰; **B**, a tertiary (3°) amide fumaramide group, should bind less well than **A** because the extra methyl groups are bulky and some of the 3° amide bond rotamers will be sterically mismatched with the benzylic amide macrocycle¹¹; **C**, a succinic amide ester, can still engage in double bifurcated hydrogen bonding, but will do so only weakly compared to **A** or **B** because it is flexible and one of the groups is an ester, which is a poor hydrogen bond acceptor^{10,12}. A fourth station, an isolated amide group (shown as **D** in *E*, **3**) that can make fewer intercomponent hydrogen bonding contacts than **A**, **B** or **C**, is also present but only affects the behaviour of the [3]catenane. A benzophenone unit was attached to the highest-affinity fumaramide group to enable selective sensitized isomerization of that station at 350 nm (ref. 11).

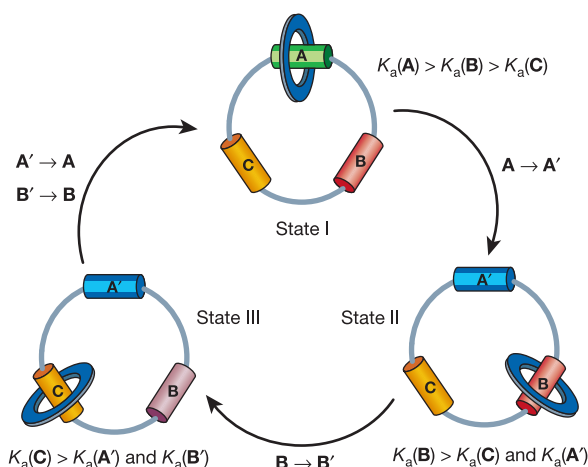


Figure 1 Stimuli-induced sequential movement of a macrocycle between three different binding sites in a [2]catenane. The larger macrocycle contains three stations, **A**, **B** and **C**, each with different binding affinities (association constants, K_a) for the smaller macrocycle such that $K_a(\mathbf{A}) > K_a(\mathbf{B}) > K_a(\mathbf{C})$. Thus, in State I, the small macrocycle preferentially resides on station **A**. If binding site **A** is converted to **A'**, a group with a lower binding affinity for the macrocycle than **B** or **C**, then the small macrocycle will move through biased brownian motion to site **B** (State II). Similarly, if **B** is then changed into a station **B'** such that $K_a(\mathbf{B}') < K_a(\mathbf{C})$ then the macrocycle will move to site **C** (State III). Finally, changing **A'** back to **A** and **B'** to **B** returns the small ring to its original position.

Studies on model [2]rotaxanes showed that (1) the benzylic amide macrocycle exhibits excellent positional discrimination between various pairs of binding sites¹², (2) the strong binding affinity of the fumaramide stations is greatly reduced upon photoisomerization to the corresponding maleamides (*Z*-olefins)¹¹, and (3) the difference in distances between the benzophenone and the two fumaramide groups in **1** is sufficient to allow complete discrimination between the sites in the sensitized photoisomerization reaction.

Interconversion of the diastereomers for each of the three structures—the macrocycle (**2**), the [2]catenane (**1**) and the [3]catenane (**3**)—was carried out in the following sequence (with reaction conditions in parentheses): *E,E* → *Z,E* (350 nm, CH₂Cl₂, 5 min, 65–67%) → *Z,Z* (254 nm, CH₂Cl₂, 20 min, 48–51%) → *E,E* (heat, 100 °C, C₂H₂Cl₄, 24 h, ~100% or catalytic ethylene diamine, 50 °C, 48 h, 50–74%). To be sure of the location of the rings after each olefin isomerization reaction, authentic samples of each diastereomer of **1–3** were isolated at each stage. Since the xylene rings of the benzylic amide macrocycle shield the part of the larger macrocycle on which it sits, its position could be unambiguously determined by comparison of the ¹H NMR spectra (Fig. 3B and Fig. 4) of the catenanes with the corresponding diastereomer of **2**. The resulting structures are shown next to each spectrum.

The ¹H NMR spectra show that the benzylic amide macrocycle in [2]catenane **1** does, indeed, move with excellent positional integrity from **A** → **B** → **C** → **A** during the chemical transformation *E,E*-**1** → *Z,E*-**1** → *Z,Z*-**1** → *E,E*-**1**. This is the first example of a catenane in which a ring can be switched between three different stations, but the rotation is not unidirectional. Although the benzylic amide macrocycle changes its position in response to the

external stimuli in discrete steps, the route it takes to get there is not directionally biased. Over the complete sequence of reactions an equal number of macrocycles go from **A** through **B** and **C** and back to **A** in each direction.

To bias the direction the macrocycle takes from station to station, barriers must exist to prevent brownian motion in a particular direction. The barriers must be transient, however, to allow 360° rotation of the macrocycle. Such a situation is inherently present in a [3]catenane if each ring is able to block the passage of the other one in a particular direction (Fig. 5 and Methods). Thus irradiation at 350 nm of *E,E*-**3** causes counter-clockwise (as drawn in Fig. 5) rotation of the blue macrocycle to the succinic amide ester (orange) station to give *Z,E*-**3**. Isomerization (254 nm) of the remaining fumaramide group causes the other (purple) macrocycle to relocate to the single amide (dark green) station (*Z,Z*-**3**) and, again, this occurs counter-clockwise because the clockwise route is blocked by the other (blue) macrocycle. This ‘follow-the-leader’ process, each macrocycle in turn moving and then blocking a direction of passage for the other macrocycle, is repeated throughout the sequence of transformations shown in Fig. 5. After three diastereomer interconversions *E,E*-**3** is again formed but 360° rotation of each of the small rings has not yet occurred, they have only swapped places. Complete unidirectional rotation of both small rings occurs only after the synthetic sequence (1)–(3) has been completed twice. Somewhat counter-intuitively, the small macrocycles in the [2]- and [3]catenanes move from station to station in opposite directions in response to the same series of chemical reactions.

Unlike the directionally rotating systems designed by Kelly^{3,4} and Feringa^{5–8}, neither **1** nor **3** are three-dimensionally chiral, but they are chiral in the two dimensions in which the rotation occurs.

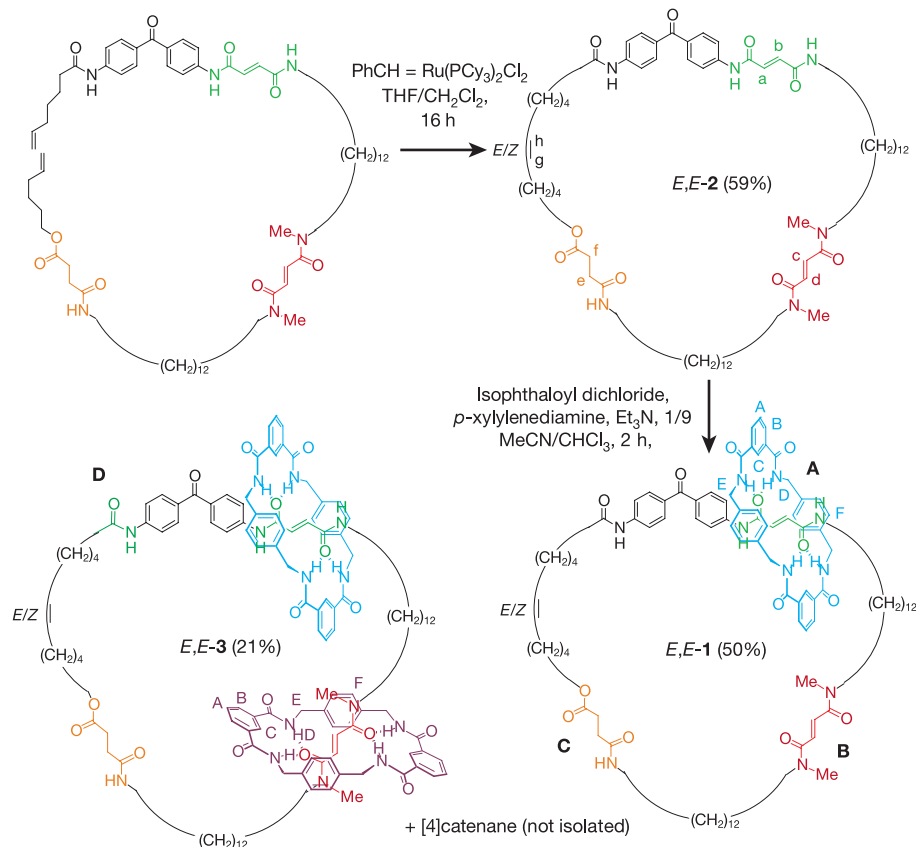


Figure 2 Ring-closing reactions to form macrocycle *E,E*-**2**, [2]catenane *E,E*-**1** and [3]catenane *E,E*-**3**. Closure of the large macrocycle, *E,E*-**2**, proceeds via ring-closing metathesis with the first-generation Grubbs catalyst PhCH = Ru(PCy₃)₂Cl₂ (ref. 19). The

remarkable yield (59%) for the formation of the 63-membered ring is probably a result of intramolecular hydrogen bonding folding the open-chain precursor. The coloured letters indicate ¹H assignments in Figs 3B and 4.

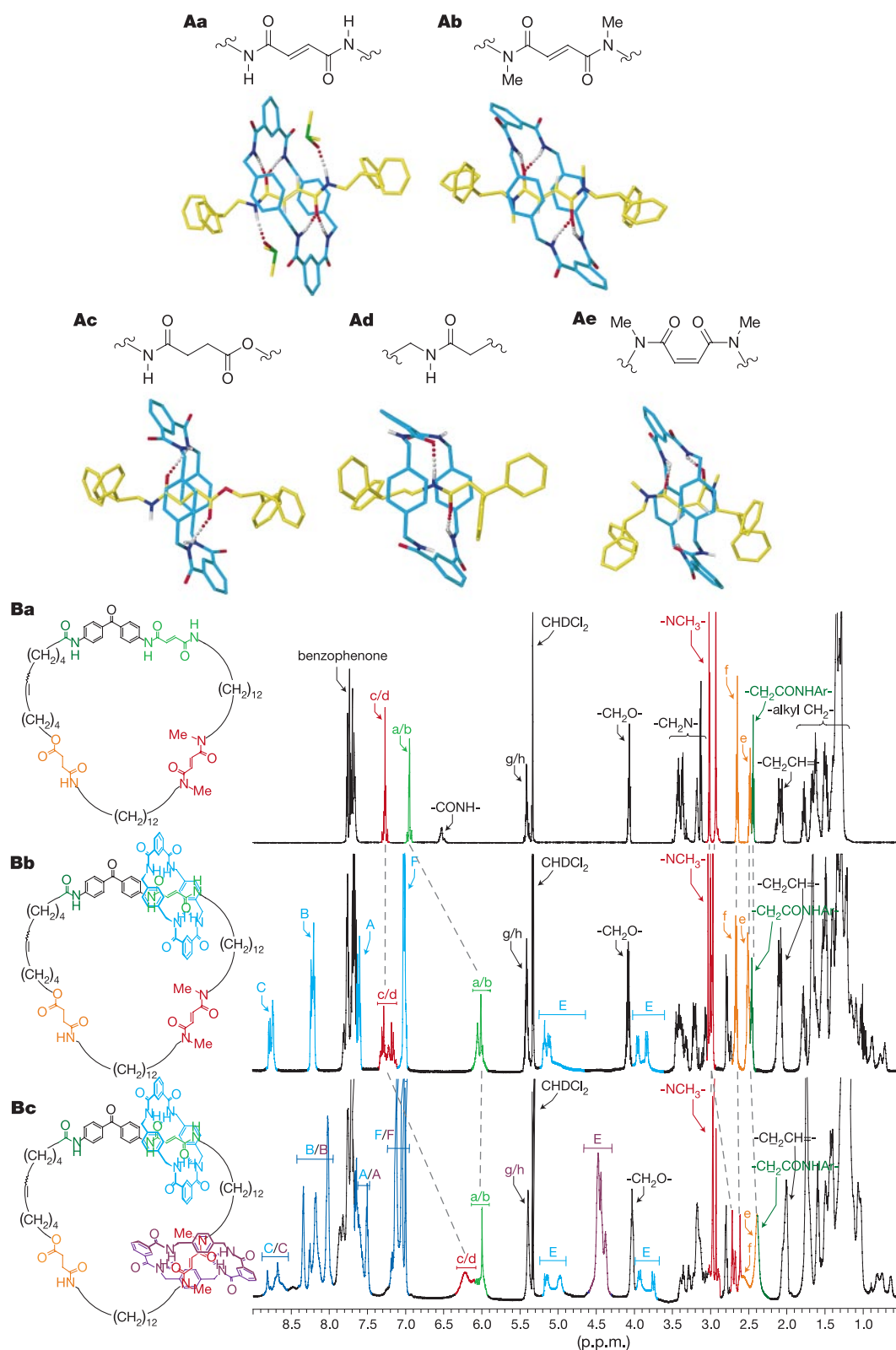


Figure 3 Structural characterization of *E,E*-1-3 and model compounds. **A**, X-ray structures of single binding site [2]rotaxanes showing solid-state hydrogen-bonding motifs of various stations in order of predicted macrocycle binding affinity. **A, a**, Fumaramide; **A, b**, *N,N'*-dimethylfumaramide; **A, c**, succinic amide ester; **A, d**, single amide; **A, e**, *N,N'*-dimethylmaleimide. **B**, 600 MHz ^1H NMR spectra (298 K, CD_2Cl_2) of the following molecules. **B, a**, Macrocycle *E,E*-2; **B, b**, [2]catenane *E,E*-1

($\Delta G^\ddagger_{\text{pirouette}}$ $14.2 \pm 0.2 \text{ kcal mol}^{-1}$); **B, c**, [3]catenane *E,E*-3 ($\Delta G^\ddagger_{\text{pirouette}}$ 14.2 ± 0.2 and $<9.7 \text{ kcal mol}^{-1}$). The chemical structures show the positions of the macrocycles indicated by the ^1H NMR chemical shifts. Activation energy barriers ($\Delta G^\ddagger_{\text{pirouette}}$) for the pirouetting of the benzylic amide macrocycles about each station were determined by the coalescence method and the rates extrapolated to 298 K.

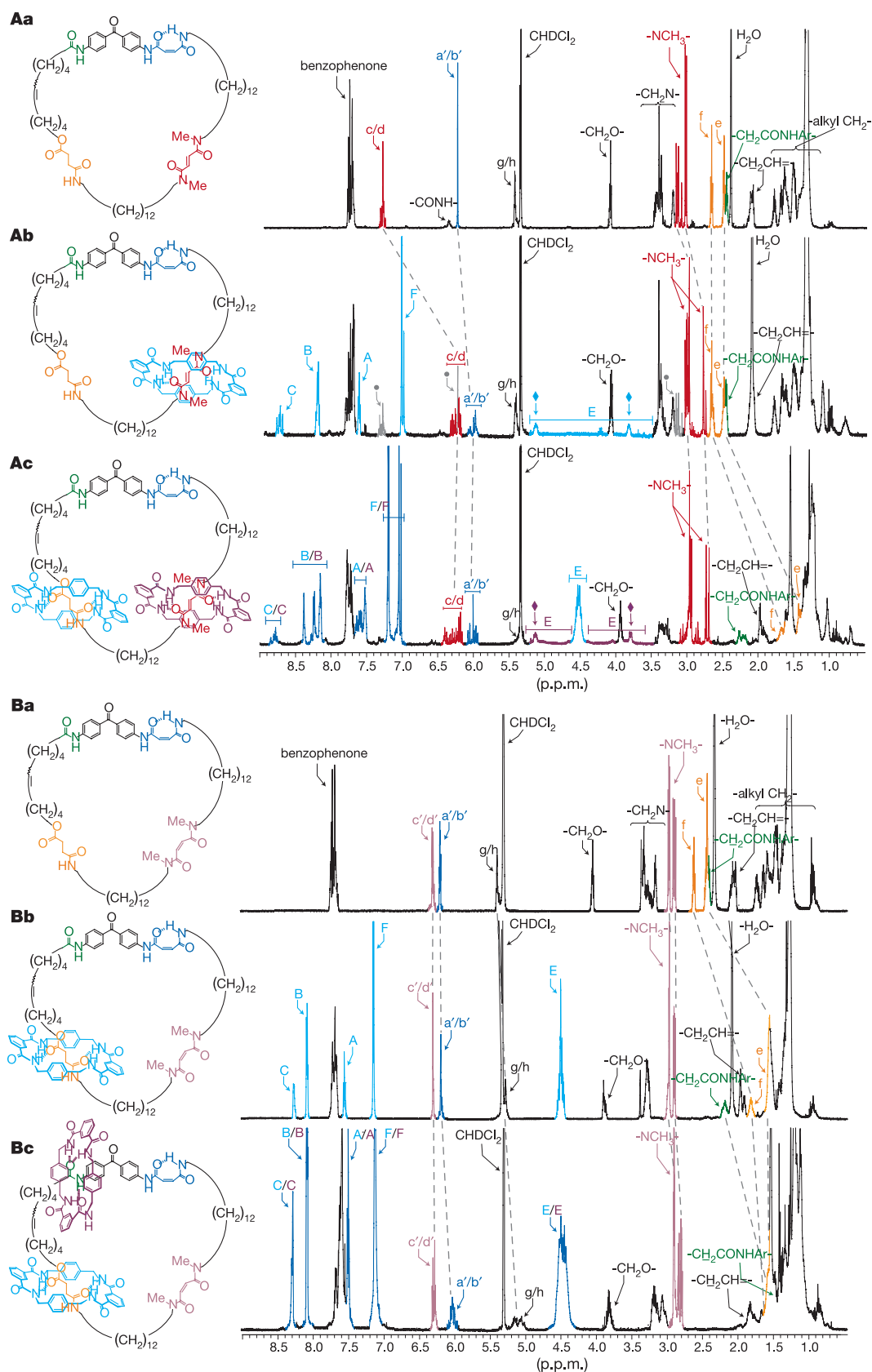


Figure 4 Structural characterization of photoisomerized macrocycles and catenanes *Z,E*-1-3 and *Z,Z*-1-3. 600 MHz ^1H NMR spectra (298 K, CD_2Cl_2). **A, a**, Macrocycle *Z,E*-2, **A, b**, [2]catenane *Z,E*-1 ($\Delta G_{\text{pirouette}}^{\ddagger} = 13.0 \pm 0.2 \text{ kcal mol}^{-1}$), **A, c**, [3]catenane *Z,E*-3 ($\Delta G_{\text{pirouette}}^{\ddagger} = 13.0 \pm 0.2$ and $<9.7 \text{ kcal mol}^{-1}$). **B, a**, Macrocycle *Z,Z*-2; **B, b**, [2]catenane *Z,Z*-1 ($\Delta G_{\text{pirouette}}^{\ddagger} < 9.7 \text{ kcal mol}^{-1}$); **B, c**, [3]catenane *Z,Z*-3 ($\Delta G_{\text{pirouette}}^{\ddagger}$ both $<9.7 \text{ kcal mol}^{-1}$). Grey indicates residual macrocycle; blue and pink

diamonds indicate minor 3° amide rotamers. The isolation of the individual macrocycle and catenane isomers was technically challenging: both *E* and *Z* isomers are present at the steady state in each photochemical reaction (the wavelengths for the reactions have not yet been optimized) requiring the separation of compounds of 1–2.5 kDa molecular weight, which differ only in the stereochemistry of one double bond.

Introducing 3D chirality into **3** would only define a location from which to describe the direction of the rotation. Such a reference point is arbitrary and does not influence unidirectional behaviour. Consider, for example, a transparent clock with symmetrical digits I, II, III, IIII..., such that it is achiral. Although the hands move unidirectionally from I $\rightarrow$ II $\rightarrow$ III and so on, to an observer in front of the clock the hands move clockwise, whereas to an observer from behind they move counter-clockwise. Exchanging the positions of the observers does not reverse the unidirectional sequence through which the clock hands move, it changes only the direction in which the observers see them move. Neither macroscopic nor chemical chirality are required to do directional work with a rotating machine. Rotation of a chiral moiety, or something achiral within a chiral environment, is only necessary to do work orthogonal to the plane of rotation (as with a propeller, for example). A winch, operating on whatever length scale, can rotate in either direction in order to wind up a chain—and if there are many winches and many chains, it is irrelevant whether all of the winches rotate in the same direction or not.

But the system does make mistakes—just as the biological rotary motor F_1 -ATPase has been observed¹³ to rotate 120° in the wrong direction every so often during its ‘unidirectional’ cycle, the small macrocycles in **3** will not move solely in the directions indicated in Fig. 5. In addition to some rings occasionally taking less-kinetically favoured pathways to the thermodynamic minima, the background thermal energy will also cause the small macrocycles to move even when no stimuli are applied and the direction of this non-biased brownian motion will be random. However, the significantly higher activation barrier for circumrotation of the large macrocycle ($\Delta G^\ddagger > 23 \text{ kcal mol}^{-1}$ in CDCl_3 at 298 K, see Supplementary Information) compared to the stimuli-induced translational barriers for

the small rings ($< 8 \text{ kcal mol}^{-1}$ in CDCl_3 at 298 K for all except $\text{C} \rightarrow \text{B}$, which is $11.3 \text{ kcal mol}^{-1}$) means the random background rotation can be minimized by lowering the temperature. The reaction sequence in Fig. 5 was repeated at -78°C using photochemically generated bromine radicals to catalyse the maleamide $\rightarrow$ fumaramide isomerization step¹⁴. At this temperature the photo-stationary states for steps (1) and (2) are still reached within 5 and 20 min, respectively, and after 10 min treatment of *Z,Z*-**3** with Br_2 (~ 1 equivalent) under 400–670 nm irradiation $\sim 100\%$ of *E,E*-**3** had been re-formed. At 298 K the frequency of background circumrotation of *E,E*-**3** is approximately once every 8 hours and the rate approximately halves with every 10°C reduction in temperature. Consequently, **3** is a rotary motor with moving parts that will slip, occasionally or frequently depending upon the reaction conditions. Actually, this is largely irrelevant to the motor’s function as the time-averaged result is always the same net rotation of the submolecular components in one direction (as long as the directionality of the stimuli-induced motion is not compromised at higher temperatures). Only fuelled molecular motions can do work^{15,16} and, as with all molecular motors, the performed work is unaffected by the degree of background thermal motion; a winch will wind up the same length of chain whether it rotates 100 times clockwise and 99 times counterclockwise or just once (in either direction).

It is interesting to consider the energetics involved in how these systems function. Indeed, there are parallels to the ways that biological motors appear to operate¹. The movement of the macrocycles from station to station in **3** is brought about by using chemical reactions to put the molecule into non-equilibrium conformations and then allowing brownian motion to drive the macrocycle to the new global minimum in the most kinetically accessible direction. The amount of energy potentially available from the light-fuelled (and heat- or chemically promoted) motions in both **1** and **3** corresponds to the energy released by non-covalent bonding as the rings move from their non-equilibrium to equilibrium positions¹⁷, that is, the difference in macrocycle-binding energies of the *A'* and *C* stations ($\Delta\Delta G_{A'-C}$) for Fig. 5 step (1), $\Delta\Delta G_{B'-D}$ for Fig. 5 step (2), $\Delta\Delta G_{D-A} + \Delta\Delta G_{C-B}$ for Fig. 5 step (3) and so on. The percentage of this energy available to do work through directional mechanical motion in **3** depends on the degree of directionality of the stimuli-induced movement. Although this may be high, even in this first-generation mechanically linked rotor, harnessing the potential of powered mechanical motions to do useful work with a synthetic molecular machine still presents a significant challenge.

The dynamic behaviour of the small macrocycles in the [2] and [3]catenanes at equilibrium is also interesting. A 180° pirouette of the small macrocycle around the axis of the large macrocycle (plus a chair–chair flip) groups the eight H_E protons into only two different sets of averaged environments and the shape of the H_E signals in the ^1H NMR spectra can thus give information regarding the dynamics of these processes. Indeed, the very different appearances of the H_E protons in Fig. 3B, b, and Fig. 4A, b and Fig. 4B, b are a result of the rate of pirouetting of the small macrocycle speeding up by over three orders of magnitude as it moves from station to station around the ring. This is perhaps not surprising given that the ring is moving from stronger binding sites to weaker ones; however, the small macrocycles in the [3]catenane do not always behave in the same way. Although the H_E protons of the major 3° amide rotamers are broad when the macrocycle occupies the methylated fumaramide station in *Z,E*-**1** (Fig. 4A, b), the H_E protons of the macrocycle at the same station in the [3]catenane *E,E*-**3** (Fig. 3B, c) are in rapid exchange indicating fast pirouetting (we cannot strictly identify which H_E signals correspond to which macrocycle in *E,E*-**3**, but this is the obvious assignment and is consistent with modelling studies). However, upon photoisomerization of *E,E*-**3** to *Z,E*-**3** the macrocycle over the methylated fumaramide station then pirouettes

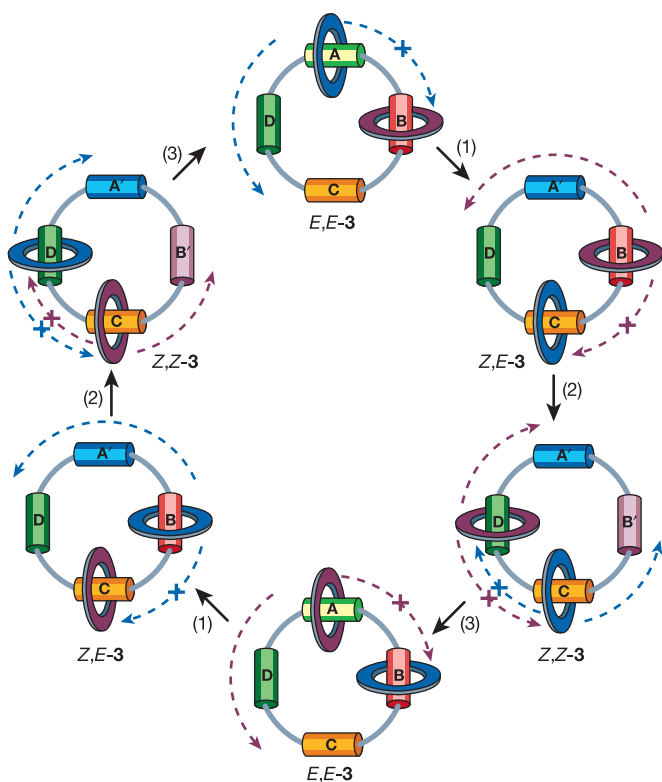


Figure 5 Stimuli-induced unidirectional rotation in a four-station [3]catenane, **3**. (1) 350 nm, CH_2Cl_2 , 5 min, 67%; (2) 254 nm, CH_2Cl_2 , 20 min, 50%; (3) heat, 100°C , $\text{C}_2\text{H}_2\text{Cl}_4$, 24 h, $\sim 100\%$; catalytic ethylenediamine, 50°C , 48 h, 65%; or catalytic Br_2 , 400–670 nm, CH_2Cl_2 , -78°C , 10 min, $\sim 100\%$.

slowly (that is, the major 3° amide rotamer H_E signals broaden in Fig. 4A, c. There is no ambiguity in the macrocycle assignment in Z,E-3 because of the visible presence of different 3° amide rotamers). Moving the macrocycle from the 2° amide fumaramide station in E,E-3 to the succinic amide ester station in Z,E-3 (during which the pirouetting frequency of that macrocycle increases) thus causes the rate of pirouetting of the other macrocycle to decrease, attributed by molecular modelling (Supplementary Information) to the different hydrogen-bonding patterns in the various [3]catenanes. Using stimuli-induced mechanical movement in one component to induce or influence dynamics in another is still rare in synthetic molecular systems¹⁸, yet it is the principle which underpins the workings of all macroscopic machines with moving parts. □

Methods

Kinetic studies on model [2]rotaxanes

The small rings in a [3]catenane will only undergo unidirectional rotation if the 'blocking' ring is effectively translationally immobile over the time period that the other ring moves. If that is not the case, then instead of a single macrocycle moving counter-clockwise during a particular step, both small rings could move clockwise (one displacing the other), or a combination of these could occur throughout the reaction sequence, removing the directionality of the motion. To show that the components of **3** undergo unidirectional rotation over a range of conditions, the contributions of these processes were estimated from the kinetics of various model compounds (Supplementary Information). The energy barriers for a benzylic amide macrocycle to move from each type of station (A and B, C, D, A' and B') to another station 12 carbon atoms away at 298 K in CDCl₃ were experimentally determined in symmetrical two-station [2]rotaxanes by variable-temperature ¹H NMR spectroscopy. The benzophenone unit was shown not to significantly slow shuttling using another model [2]rotaxane. The experimentally determined barriers (fumaramide and *bis*-N-methyl fumaramide 16.2 ± 0.4 kcal mol⁻¹; succinic amide ester 11.3 ± 0.2 kcal mol⁻¹; amide < 8 kcal mol⁻¹; maleamide and *bis*-N-methyl maleamide < 8 kcal mol⁻¹) mean that at 298 K in CDCl₃ stations A and B decompose, and the macrocycle moves to the next station, 4,000 times less frequently than C and >10⁶ times less frequently than D, A' and B' (the ratio of rates is given by $e^{\Delta\Delta G^\ddagger/k_B T}$). Thus, starting from E,E-3, when A is isomerized to A' the macrocycle initially at A' will move to C (overcoming two barriers of less than 8 kcal mol⁻¹) to arrive at the equilibrium position of Z,E-3 at least a million times more often than the macrocycle originally at B moving to C (barrier of 16.2 kcal mol⁻¹) and the macrocycle originally at A moving to B. Similarly, movement of the macrocycle from the B' station to D to give the most stable positional isomer of Z,Z-3 will occur three orders of magnitude faster than macrocycles moving from C to D and B' to C. Finally, as long as the N-methyl maleamide station is not isomerized to the fumaramide unit (B' → B) at a significantly faster rate than the secondary maleamide station (A' → A) as Z,Z-3 is converted to E,E-3 (the rates we observe experimentally are the same for both stations), then the macrocycle initially at D will move rapidly to A where it will be bound tightly and therefore be thousands of times slower to move to the vacant B station than the macrocycle originally at C. While it is always possible that the inter-station dynamics in the model compounds differ to those in the [3]catenane, the kinetics suggest overwhelming directionality for the stimuli-induced motion in **3** over a wide range of conditions.

Received 19 February; accepted 21 May 2003; doi:10.1038/nature01758.

- Schliwa, M. (ed.) *Molecular Motors* (VCH-Wiley, Weinheim, 2003).
- Balzani, V., Credi, A., Raymo, F. M. & Stoddart, J. F. Artificial molecular machines. *Angew. Chem. Int. Edn* **39**, 3348–3391 (2000).
- Kelly, T. R., De Silva, H. & Silva, R. A. Unidirectional rotary motion in a molecular system. *Nature* **401**, 150–152 (1999).
- Kelly, T. R., Silva, R. A., De Silva, H., Jasmin, S. & Zhao, Y. A rationally designed prototype of a molecular motor. *J. Am. Chem. Soc.* **122**, 6935–6949 (2000).
- Koumura, N., Zijlstra, R. W. J., van Delden, R. A., Harada, N. & Feringa, B. L. Light-driven monodirectional molecular rotor. *Nature* **401**, 152–155 (1999).
- Feringa, B. L., Koumura, N., van Delden, R. A. & ter Wiel, M. K. J. Light-driven molecular switches and motors. *Appl. Phys. A* **75**, 301–308 (2002).
- Koumura, N. L., Geertsema, E. M., van Gelder, M. B., Meetsma, A. & Feringa, B. L. Second generation light-driven molecular motors. Unidirectional rotation controlled by a single stereogenic center with near-perfect photoequilibria and acceleration of the speed of rotation by structural modification. *J. Am. Chem. Soc.* **124**, 5037–5051 (2002).
- van Delden, R. A., Koumura, N., Harada, N. & Feringa, B. L. Unidirectional rotary motion in a liquid crystalline environment: Color tuning by a molecular motor. *Proc. Natl Acad. Sci. USA* **99**, 4945–4949 (2002).
- Sauvage, J.-P. & Dietrich-Buchecker, C. (eds) *Molecular Catenanes, Rotaxanes and Knots* (VCH-Wiley, Weinheim, 1999).
- Gatti, F. G. et al. Stiff, and sticky in the right places: The dramatic influence of preorganizing guest binding sites on the hydrogen bond-directed assembly of rotaxanes. *J. Am. Chem. Soc.* **123**, 5983–5989 (2001).
- Gatti, F. G. et al. Photoisomerization of a rotaxane hydrogen bonding template: Light-induced acceleration of a large amplitude rotational motion. *Proc. Natl Acad. Sci. USA* **100**, 10–14 (2003).
- Altieri, A. et al. Remarkable positional discrimination in bistable, light and heat-switchable, hydrogen bonded molecular shuttles. *Angew. Chem. Int. Edn* **42**, 2296–2300 (2003).

- Yasuda, R., Noji, H., Kinoshita, K. Jr & Yoshida, M. F₁-ATPase is a highly efficient molecular motor that rotates with discrete 120° steps. *Cell* **93**, 1117–1124 (1998).
- Frkanec, L., Jokić, M., Makarević, J., Wolsperger, K. & Žinić, M. Bis(PheOH) maleic acid amide-fumaric acid amide photoisomerization induces microsphere-to-gel fibre morphological transition: A photoinduced gelation system. *J. Am. Chem. Soc.* **124**, 9716–9717 (2002).
- Feynman, R., Leighton, R. B. & Sands, M. L. *The Feynman Lectures on Physics* Vol. 1, Ch. 46 (Addison-Wesley, Reading, MA, 1963).
- Davis, A. P. Tilting at windmills? The second law survives. *Angew. Chem. Int. Edn* **37**, 909–910 (1998).
- Brouwer, A. M. et al. Photoinduction of fast, reversible translational motion in a hydrogen-bonded molecular shuttle. *Science* **291**, 2124–2128 (2001).
- Kelly, T. R. et al. A molecular brake. *J. Am. Chem. Soc.* **116**, 3657–3658 (1994).
- Trnka, T. M. & Grubbs, R. H. The development of L₂X₂Ru = CHR olefin metathesis catalysts: An organometallic success story. *Acc. Chem. Res.* **34**, 18–29 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank E. M. Perez for developing the catalytic Br₂ reaction conditions and P. J. Nash for several of the model [2]rotaxanes. This work was carried out through the support of the RTN network EMMMA and the FET MechMol. D.A.L. is an EPSRC Advanced Research Fellow.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.A.L. (David.Leigh@ed.ac.uk).

Post-earthquake ground movements correlated to pore-pressure transients

Sigurjón Jónsson*, Paul Segall†, Rikke Pedersen‡ & Grímur Björnsson§

* Department of Earth and Planetary Sciences, Harvard University, 20 Oxford Street, Cambridge, Massachusetts 02138, USA

† Department of Geophysics, Stanford University, Stanford, California 94305-2215, USA

‡ Nordic Volcanological Institute, Grensásvegur 50, and § National Energy Authority, Grensásvegur 9, 108 Reykjavík, Iceland

Large earthquakes alter the stress in the surrounding crust, leading to triggered earthquakes and aftershocks^{1–3}. A number of time-dependent processes, including afterslip, pore-fluid flow and viscous relaxation of the lower crust and upper mantle, further modify the stress and pore pressure near the fault, and hence the tendency for triggered earthquakes^{4,5}. It has proved difficult, however, to distinguish between these processes on the basis of direct field observations, despite considerable effort⁶. Here we present a unique combination of measurements consisting of satellite radar interferograms⁷ and water-level changes in geothermal wells following two magnitude-6.5 earthquakes in the south Iceland seismic zone. The deformation recorded in the interferograms cannot be explained by either afterslip or viscoelastic relaxation, but is consistent with rebound of a porous elastic material in the first 1–2 months following the earthquakes. This interpretation is confirmed by direct measurements which show rapid (1–2-month) recovery of the earthquake-induced water-level changes. In contrast, the duration of the aftershock sequence is projected to be ~3.5 years, suggesting that pore-fluid flow does not control aftershock duration. But because the surface strains are dominated by pore-pressure changes in the shallow crust, we cannot rule out a longer pore-pressure transient at the depth of the aftershocks. The aftershock duration is consistent with models of seismicity rate variations based on rate- and state-dependent friction laws.

The south Iceland seismic zone (SISZ) is a left-lateral transform

zone connecting two sections of the mid-Atlantic plate boundary (Fig. 1), where east–west transform motion is accommodated by north–south-oriented right-lateral faults⁸. This region has suffered more than 30 large earthquakes during the past 800 years⁹. The earthquakes on 17 and 21 June 2000 ruptured two of these faults, separated by 17 km (refs 10–12) (Fig. 1). Large earthquakes in the SISZ typically cluster in time; five or six large earthquakes occurred during a period of only two weeks in 1896 (ref. 13). The last major earthquake occurred in 1912, and the time interval between sequences is 45–112 years (ref. 13).

Satellite radar interferograms (InSAR)⁷ of the SISZ exhibit a quadrantal pattern of post-seismic deformation following the June 2000 earthquakes. The interferograms show several centimetres of LOS (line of sight between the ground and the satellite radar) displacement that is opposite in sense to that from right-lateral coseismic fault slip (Fig. 2 and ref. 11). An interferogram spanning the time interval from 19 June 2000 (between the two earthquakes) to 24 July 2000 shows localized deformation near the eastern, 17 June rupture (Fig. 2a). Following the earthquakes, the ground moved up towards the satellite radar in quadrants of coseismic extension (northeast and southwest of the fault) and away in quadrants of coseismic compression (northwest and southeast of the fault). The largest LOS displacement is observed northwest of the fault (–3.5 cm), while the two quadrants to the south exhibit smaller signals. Two other post-seismic interferograms exhibit similar deformation near the 17 June rupture (Fig. 2f, g). These interferograms also show LOS ground displacements up towards (away from) the satellite radar, northeast (northwest) of the 21 June rupture, but only LOS displacements away from the radar near the southern end of the fault. A fourth interferogram, spanning 35 days from 12 August to 16 September shows little or no deformation

(Fig. 2e). Therefore, the post-seismic deformation was a transient that lasted about 2 months.

Post-seismic deformation can be caused by: (1) creep (afterslip) on or adjacent to the mainshock rupture¹⁴; (2) visco-elastic relaxation in the lower crust and upper mantle driven by stresses induced by the mainshock⁶; (3) poro-elastic rebound due to pore-fluid flow in response to mainshock induced pore-pressure changes¹⁵. We show here that the SISZ post-seismic deformation is consistent with poro-elastic rebound, but inconsistent with both afterslip and visco-elastic relaxation.

The observed post-seismic LOS displacements are opposite in sense to those from right-lateral strike slip. If the observed signals were due to creep, as initially was suggested¹⁰, the slip would require left-lateral backslip with a significant component of vertical slip. The latter is unlikely on these near-vertical faults because no vertical slip occurred coseismically¹¹. Although the 21 June earthquake altered the stress on the 17 June fault, these changes do not favour backslip, as the calculated Coulomb failure stress changes have the wrong sign to encourage such slip.

Viscous flow of the lower crust or upper mantle is also not a plausible explanation. The short duration of the deformation transient would require a viscosity of the order of 10^{17} Pa s (relaxation time of 2–3 weeks), which is much lower than viscosity estimates for the lower crust and upper mantle in Iceland; these estimates range from 1×10^{18} to 5×10^{19} Pa s (refs 16–19). Even though the viscosity may be low initially after a large stress change⁶, a lower-crustal or upper-mantle origin of the signals cannot be reconciled with the limited spatial extent of the observed deformation, which is concentrated within 5–10 km of the faults (Fig. 2a, d).

Large earthquakes cause pore-pressure increase in areas of

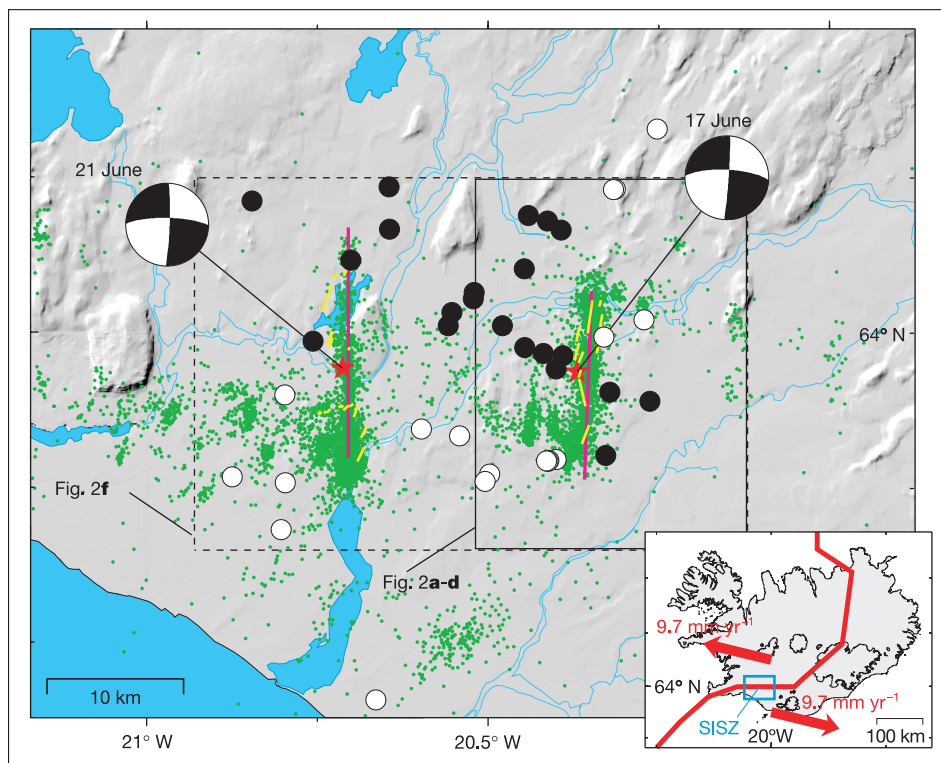


Figure 1 Map showing the location of the two June 2000 (moment magnitude $M_w = 6.5$) earthquakes in south Iceland. Inset, the approximate location of the mid-Atlantic plate boundary in Iceland, which has a full spreading rate of 1.94 cm yr^{-1} (ref. 25). The south Iceland seismic zone (SISZ) is an east–west-oriented left-lateral transform zone (small blue box marks area shown in detail). The map shows mainshock epicentres (red stars) and focal mechanisms (Harvard CMT²⁰), other earthquakes from 21 June to 31 December

2000 (green dots), mapped surface ruptures²² (yellow lines), and model fault traces as estimated from an inversion of the coseismic GPS and InSAR data for fault geometry¹¹ (purple lines). Also shown is the 17 June coseismic water-level increase (black dots) and decrease (white dots) in geothermal wells in the area²⁰. Note water-level increase in compressional quadrants. Solid rectangle marks the area shown in Fig. 2a–d and the larger dashed rectangle the area covered in Fig. 2f.

compression and decrease in areas of dilation^{5,20}. These pore-pressure gradients induce groundwater flow and additional time-dependent strain²¹. This poro-elastic rebound results in subsidence within (coseismic) compressional quadrants and uplift in (coseismic) extensional quadrants. Assuming complete equilibration of the pore-pressure gradients yields a quadrantal deformation pattern similar to the observations, especially near the 17 June rupture (Fig. 2b). Although modelled displacements capture the main features of the interferometric signals, some differences are evident. Observed deformation near the southern end of the 17 June rupture is weaker than at the northern end, indicating either spatially variable material properties or local rebound that occurred so rapidly it was not detected. The southern end of the 21 June rupture, an area of very complex surface faulting, exhibits only subsidence (Fig. 2f). Complications here include local conjugate east–west-oriented left-lateral faulting²². In contrast, the poro-elastic computations are based on a single vertical north–south fault.

Poro-elastic rebound has previously been inferred from InSAR measurements following the 1992 Landers earthquake¹⁵. We find direct confirmation of pore-pressure recovery using water-level changes observed in numerous geothermal wells in the SISZ. The sign of the coseismic water-level changes exhibits a quadrantal pattern, consistent with the predicted undrained response (Figs 1 and 3a). Water-level changes reverse sign in the post-seismic period (Fig. 3b). Although there is some variability between wells, the

duration of the water-level recovery is typically 1–2 months, consistent with the duration of the deformation transient (Fig. 4a). Interestingly, the two wells located just south of the 17 June fault show faster water-level recovery, indicating high permeability²⁰ (see well HR in Fig. 4a). This suggests that the weak deformation signal observed in this area resulted from a fast rebound that was not captured by the interferograms. The initial images were acquired 2 days (Fig. 2a) and 7 days (Fig. 2f) after the 17 June earthquake.

Competing models for the time-dependent decay of aftershocks include changes in fault strength due to poro-elastic relaxation⁵, and delayed slip instabilities due to rate- and state-dependent friction on faults stressed by the mainshock². We now consider if the unique measurements collected following the June 2000 earthquakes allow us to discriminate between these models.

Off-fault aftershocks preferentially occur in quadrants that experienced decreases in pore-pressure and fault-normal extension on north–south-oriented faults, that is, northeast and southwest of the faults (Fig. 1). Post-seismic pore-fluid flow would raise pore pressure in those quadrants, bringing the faults closer to failure. However, the duration of the aftershock sequence is much longer (~3.5 years) than the duration of the observed pore-pressure transients (Fig. 4b), suggesting that pore-pressure recovery is not the controlling mechanism. On the other hand, the permeability at the 3–10-km depth of the aftershocks is likely to be less than that at

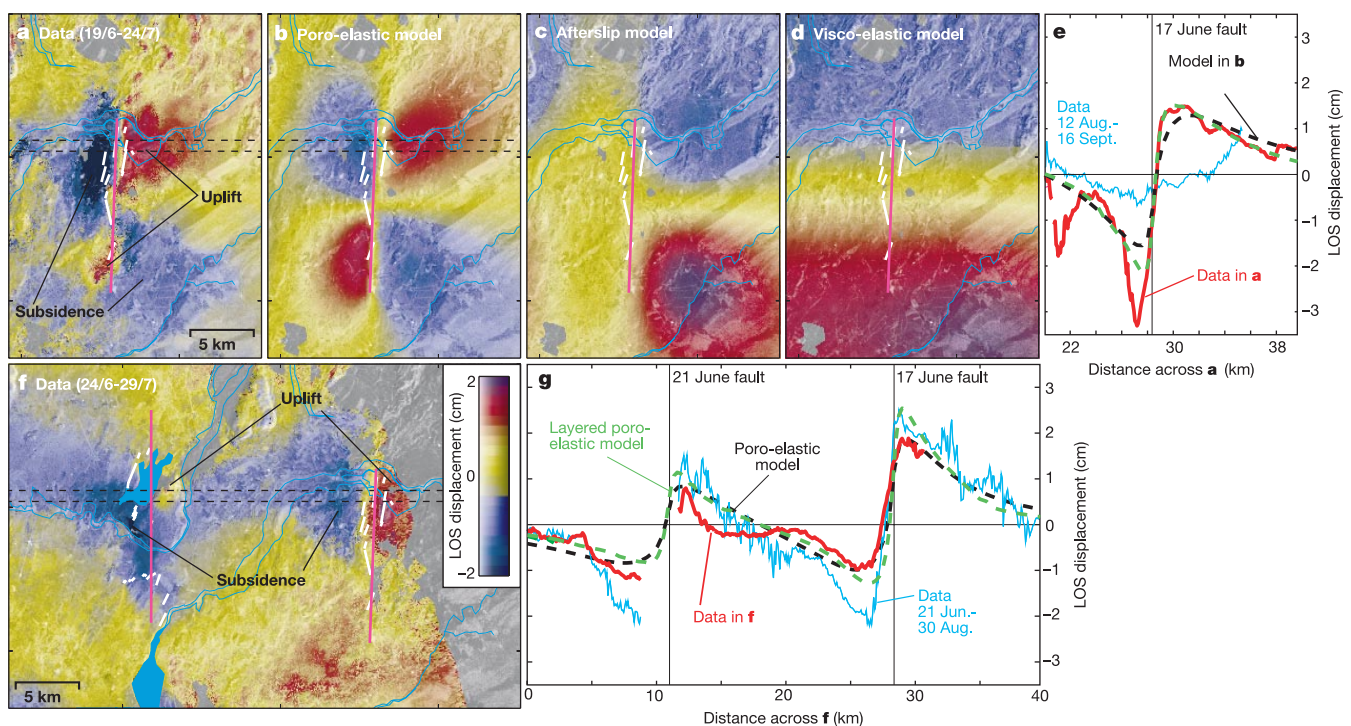


Figure 2 Synthetic aperture radar interferograms (InSAR) showing observed and simulated post-seismic deformation in south Iceland. The interferograms are unwrapped²⁷, and show ground displacements in the line of sight (LOS) towards the European radar satellite ERS-2 (range decrease is shown positive). The measurement is most sensitive to vertical ground motion, as the LOS vectors are roughly [east, north, up] = [0.38, −0.11, 0.92] for image **a** acquired from a descending orbit and [−0.41, −0.11, 0.90] for image **f** from an ascending orbit. The post-seismic interferograms span **a**, 19 June to 24 July (track 95, frame 2313) and **f**, 24 June to 29 July (track 173, frame 1287). In **a**, coseismic deformation caused by the 21 June earthquake, 17 km to the west, has been removed using a fault-slip model¹¹. Also shown are simulated interferograms of post-seismic deformation using poro-elastic (**b**), right-lateral afterslip (**c**), and visco-elastic (**d**) models. The poro-elastic model prediction, **b**, is calculated using an undrained Poisson's ratio ν_u of 0.31 and a drained Poisson's ratio ν of 0.27 (refs 15, 24, 28), as

well as the coseismic slip distribution for the two earthquakes¹¹. The afterslip model assumes up to 50 cm of right-lateral slip centred at 8 km depth. The visco-elastic model has a 10-km-thick elastic crust overlying a visco-elastic half-space with viscosity of 10^{17} Pa s, chosen to match the time duration of the observed deformation transient. In **e** and **g** the average of LOS displacements profiles (between dashed lines) for interferograms in **a** and **f** are shown, as well as for two other interferograms (not shown) spanning the time intervals from 21 June (22 h after the 21 June earthquake) to 30 August (ascending track 130, frame 1287), and 12 August to 16 September (descending track 367, frame 2313). The response from a layered poro-elastic model (green dashed line), with drainage in only the top 1.5 km of the crust (see Methods section), shows similar response as the homogenous model (black dashed line and **b**) when the difference of the Poisson's ratios ($\nu_u - \nu$) is scaled (here by 2).

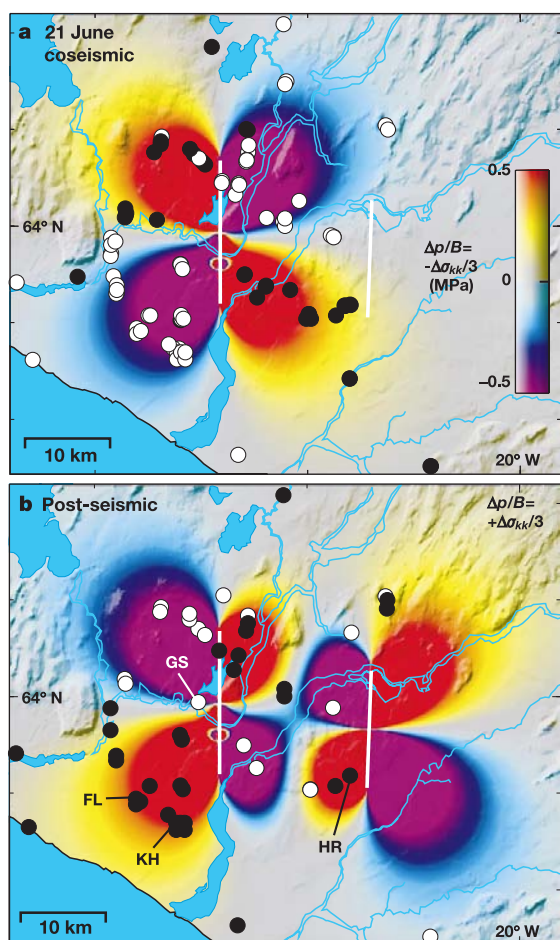


Figure 3 Coseismic and post-seismic water-level changes in geothermal wells in south Iceland. **a**, The 21 June coseismic water-level increase (black dots) and decrease (white dots) as well as the predicted coseismic pore-pressure change at 0.5 km depth (normalized by B : $\Delta p/B = -\Delta\sigma_{kk}/3$). **b**, Post-seismic water-level changes after 21 June and predicted post-seismic pore-pressure changes ($\Delta p/B = +\Delta\sigma_{kk}/3$). Water-level changes in labelled wells are shown as functions of time in Fig. 4a.

the <1.5 -km depth of the geothermal wells. To test whether the observed surface displacements might have been caused by shallow drainage, and would thus permit pore pressures at depth to decay on a timescale commensurate with the aftershocks, we computed the surface deformation for an end-member layered model. In this calculation, the permeability between the surface and depth D is infinite, and the earthquake-induced pore pressures relax instantaneously. Below D , the permeability is zero and there is no drainage. We computed the predicted range change due to complete draining of the highly permeable near-surface rocks by integrating the point source response from the surface to depth D (see Methods). Figure 2e and g shows that by adjusting the difference between undrained and drained Poisson's ratios ($\nu_u - \nu$), it is possible to fit the observations with only the shallow rocks draining. Thus, at this point, we cannot rule out deep pore-pressure changes on the timescale of the aftershocks.

Mainshock stress changes accelerate nucleation patches on neighbouring faults to instability. It has been shown that the time-to-instability relations, for fault patches subject to rate- and state-dependent friction, yield Omori-like aftershock decay where earthquake rate decreases as $1/\text{time}$ (ref. 2). This model predicts that aftershock duration t_a is proportional to $A\sigma$, and inversely proportional to the background stressing rate. Here A is an experimentally determined fault constitutive parameter (0.005–0.02 for most

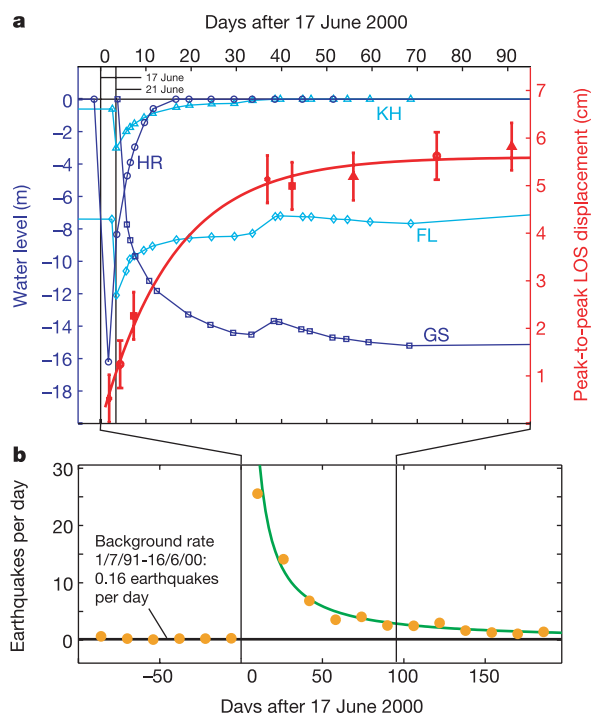


Figure 4 Water-level recoveries and aftershock decay in south Iceland. **a**, Observed peak-to-peak LOS displacement (red symbols) across the 17 June fault in several interferograms indicates that the post-seismic deformation transient lasted 2 months. Water-level changes (blue lines) show similar recovery of 1–2 months (well locations are shown in Fig. 3b). Well GS (squares) is located in coseismic compressional quadrants and experiences post-seismic water-level decline, whereas wells KH (triangles), FL (diamonds) and HR (circles) are located in coseismic extensional quadrants and exhibit post-seismic water-level increase. Note the rapid water-level recovery of well HR, located just south of the 17 June fault. **b**, Aftershock decay after the 17 June earthquake showing much longer timescale. Extrapolation of the ongoing aftershock sequence within 5 km of the 17 June fault predicts a total duration of 3.5 years. Same analysis for the 21 June fault-zone predicts aftershock duration of 3.3 years.

rocks) and σ is effective normal stress. Assuming that the background stressing rate can be approximated by earthquake stress drop $\Delta\tau$ divided by average recurrence interval t_r leads to $t_a = t_r A\sigma / \Delta\tau$. For the SISZ, where similar events recur roughly every 100 years, this model predicts an aftershock duration of 2.5–10 years (for $\Delta\tau/A\sigma$ of 10–40), which is in good accord with the observed aftershock decay.

We conclude that poro-elastic rebound dominates the post-seismic deformation in the first few months after the earthquakes. Pore pressures in the shallow crust equilibrate too rapidly to explain the observed aftershock decay, but we cannot at present rule out longer-term pore-pressure transients in lower-permeability rocks at the depths of most aftershocks. The data are consistent with aftershock duration being controlled by time-dependent failure of faults stressed by the mainshock. □

Methods

Here we explain how surface displacements due to complete draining of a permeable surface layer are calculated. Pore-pressure changes Δp induce surface displacements u_i given by

$$u_i(x) = \frac{3(\nu_u - \nu)}{B(1 + \nu_u)(1 - 2\nu)} \int \Delta p(\xi) \frac{\partial g_i^k(x, \xi)}{\partial \xi_k} dV_\xi \quad (1)$$

where ν and ν_u are the drained and undrained Poisson's ratios, B is Skempton's pore-pressure coefficient, and g_i^k is the elastic Green's function—that is, the displacement at x in the i -direction due to a point force in the k -direction at ξ (refs 23, 24). The pore-pressure change induced by dissipation of the undrained pore pressures is $B\Delta\sigma_{kk}/3$, where $\Delta\sigma_{kk}/3$ is

the coseismic mean stress change. For a homogeneous half-space we have

$$\frac{\partial g_k^k(x, \xi)}{\partial \xi_k} = \frac{(1 - 2\nu)x_i - \xi_i}{2\pi\mu R^3} \quad (2)$$

where μ is shear modulus, and R is the euclidean distance between x and ξ . Therefore, equation (1) can be written as:

$$u_i(x) = \frac{(\nu_u - \nu)}{2\pi\mu(1 + \nu_u)} \int \frac{\Delta\sigma_{kk}(\xi) x_i - \xi_i}{R^3} dV_\xi. \quad (3)$$

To determine surface displacements due to complete draining of a permeable surface layer with thickness D , overlying impermeable rocks, we integrate equation (3) from the surface to depth D . Note that equation (3) scales with $(\nu_u - \nu)$.

Received 23 January; accepted 23 May 2003; doi:10.1038/nature01776.

- Harris, R. A., Simpson, R. W. & Reasenberg, P. A. Influence of static stress changes on earthquake locations in southern California. *Nature* **375**, 221–224 (1995).
- Dieterich, J. A constitutive law for rate of earthquake production and its application to earthquake clustering. *J. Geophys. Res.* **99**, 2601–2618 (1994).
- Toda, S., Stein, R. S. & Sagiya, T. Evidence from the AD 2000 Izu islands earthquake swarm that stressing rate governs seismicity. *Nature* **419**, 58–61 (2002).
- Freed, A. M. & Lin, J. Delayed triggering of the 1999 Hector Mine earthquake by viscoelastic stress transfer. *Nature* **411**, 180–183 (2001).
- Nur, A. & Booker, J. R. Aftershocks caused by pore fluid flow? *Science* **175**, 885–887 (1972).
- Pollitz, F. F., Wicks, C. & Thatcher, W. Mantle flow beneath a continental strike-slip fault: Post-seismic deformation after the 1999 Hector Mine earthquake. *Science* **293**, 1814–1818 (2001).
- Zebker, H., Rosen, P., Goldstein, R., Gabriel, A. & Werner, C. On the derivation of coseismic displacement fields using differential radar interferometry: The Landers earthquake. *J. Geophys. Res.* **99**, 19617–19643 (1994).
- Einarsson, P. Earthquakes and present-day tectonism in Iceland. *Tectonophysics* **189**, 261–279 (1991).
- Einarsson, P. & Eiriksson, J. Earthquake fractures in the Districts Land and Rangárvellir in the South Iceland Seismic Zone. *Jökull* **32**, 113–119 (1982).
- Pedersen, R., Sigmundsson, F., Feigl, K. L. & Árnadóttir, T. Coseismic interferograms of two $M_s = 6.6$ earthquakes in the South Iceland Seismic Zone, June 2000. *Geophys. Res. Lett.* **28**, 3341–3344 (2001).
- Pedersen, R., Jónsson, S., Árnadóttir, T., Sigmundsson, F. & Feigl, K. Fault slip distribution of two June 2000 $M_w = 6.5$ earthquakes in South Iceland estimated from joint inversion of InSAR and GPS measurements. *Earth Planet. Sci. Lett.* (in the press).
- Árnadóttir, T. *et al.* Crustal deformation measured by GPS in the South Iceland Seismic Zone due to two large earthquakes in June 2000. *Geophys. Res. Lett.* **28**, 4031–4033 (2001).
- Einarsson, P., Björnsson, S., Foulger, G., Stefánsson, R. & Skaftadóttir, Th. in *Earthquake Prediction—An International Review* Vol. 4 (eds Simpson, D. W. & Richards, P. G.) 141–151 (American Geophysical Union, Washington DC, 1981).
- Bürgmann, R. *et al.* Time-space variable afterslip on and deep below the Izmit earthquake rupture. *Bull. Seismol. Soc. Am.* **92**, 126–137 (2002).
- Peltzer, G., Rosen, P., Rogez, F. & Hudnut, K. Poro-elastic rebound along the Landers 1992 earthquake surface rupture. *J. Geophys. Res.* **103**, 30131–30145 (1998).
- Sigmundsson, F. Post-glacial rebound and asthenosphere viscosity in Iceland. *Geophys. Res. Lett.* **18**, 1131–1134 (1991).
- Sigmundsson, F. & Einarsson, P. Glacio-isostatic crustal movements caused by historical volume change of the Vatnajökull ice cap, Iceland. *Geophys. Res. Lett.* **19**, 2123–2126 (1992).
- Pollitz, F. F. & Sacks, I. S. Viscosity structure beneath northeast Iceland. *J. Geophys. Res.* **101**, 17771–17793 (1996).
- Hofton, M. & Foulger, G. R. Poststrifing anelastic deformation around the spreading plate boundary, north Iceland: 1. Modeling of the 1987–1992 deformation field using a visco-elastic earth structure. *J. Geophys. Res.* **101**, 25403–25421 (1996).
- Björnsson, G., Flovenz, Ó. G., Saemundsson, K. & Einarsson, E. H. in *Proc. 26th Workshop on Geothermal Reservoir Engineering* SGP-TR-168, 327–334 (Stanford University, Stanford, 2001).
- Roeleffs, E. Poroelastic techniques in the study of earthquake-related hydrologic phenomena. *Adv. Geophys.* **27**, 135–195 (1996).
- Clifton, A. & Einarsson, P. in *Autumn Meeting 2000—Abstracts* (ed Jónsson, S. S.) 1 (Geoscientific Society of Iceland, Reykjavik, 2000).
- Segall, P. Induced stresses due to fluid extraction from axisymmetric reservoirs. *Pure Appl. Geophys.* **139**, 535–560 (1992).
- Rice, J. R. & Cleary, M. P. Some basic stress diffusion solutions for fluid-saturated elastic porous media with compressible constituents. *Rev. Geophys. Space Sci.* **14**, 227–241 (1976).
- DeMets, C., Gordon, R. G., Argus, D. F. & Stein, S. Current plate motions. *Geophys. J. Int.* **101**, 425–478 (1990).
- Dziewonski, A. M., Ekström, G. & Maternovskaya, N. Centroid-moment tensor solutions for April–June 2000. *Phys. Earth Planet. Inter.* **123**, 1–14 (2001).
- Chen, C. W. & Zebker, H. A. Two-dimensional phase unwrapping with use of statistical models for cost functions in nonlinear optimization. *J. Opt. Soc. Am.* **18**, 338–351 (2001).
- Wang, H. F. *Theory of Linear Poro-Elasticity with Applications to Geomechanics and Hydrogeology* (Princeton Univ. Press, Princeton, 2001).

Acknowledgements We thank the European Space Agency for providing the SAR data. We also thank G. Guðmundsson and R. Stefánsson for providing preliminary earthquake locations from the South Iceland Lowland (SIL) seismic network; A. Clifton and P. Einarsson for providing data of the mapped surface ruptures; F. Sigmundsson, T. Árnadóttir, K. Ágústsson, E. Roeleffs and K. Feigl for discussions; and R. Bürgmann for comments and suggestions that improved the paper.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.J. (sj@eps.harvard.edu).

Urbanization effects on tree growth in the vicinity of New York City

Jillian W. Gregg^{*‡}, Clive G. Jones[†] & Todd E. Dawson^{*‡}

^{*} Ecology and Evolutionary Biology, Cornell University, Ithaca, New York 14853, USA

[†] Institute of Ecosystem Studies, Millbrook, New York 12545, USA

Plants in urban ecosystems are exposed to many pollutants and higher temperatures, CO₂ and nitrogen deposition than plants in rural areas^{1–5}. Although each factor has a detrimental or beneficial influence on plant growth⁶, the net effect of all factors and the key driving variables are unknown. We grew the same cottonwood clone in urban and rural sites and found that urban plant biomass was double that of rural sites. Using soil transplants, nutrient budgets, chamber experiments and multiple regression analyses, we show that soils, temperature, CO₂, nutrient deposition, urban air pollutants and microclimatic variables could not account for increased growth in the city. Rather, higher rural ozone (O₃) exposures reduced growth at rural sites. Urban precursors fuel the reactions of O₃ formation, but NO_x scavenging reactions⁷ resulted in lower cumulative urban O₃ exposures compared to agricultural and forested sites throughout the northeastern USA. Our study shows the overriding effect of O₃ despite a diversity of altered environmental factors, reveals ‘footprints’ of lower cumulative urban O₃ exposures amidst a background of higher regional exposures, and shows a greater adverse effect of urban pollutant emissions beyond the urban core.

Urbanization of the globe is accelerating, with potentially large impacts on vegetation in cities and surrounding areas⁸. Urban air contains high concentrations of many gaseous, particulate and photochemical pollutants (such as NO_x, HNO₃, SO₂, H₂SO₄, O₃ and volatile organic compounds)^{1,5,6}; and urban soils are high in heavy metals and can be more hydrophobic and acidic than surrounding rural environments². Although many of these contaminants have detrimental effects on plant growth, urban environments also have higher rates of nutrient and base-cation deposition^{1,5}, warmer temperatures (urban ‘heat-island’ effect)³ and increased CO₂ concentrations⁴—factors that often, but not invariably, enhance plant growth. Given the potential for interactions among all factors⁹ and the relative absence of studies examining more than two or three factors in combination, understanding the net effect of multiple anthropogenic environmental changes in an urban environment and the relative importance of the individual factors remains a major challenge.

We used an inherently fast-growing clone of Eastern cottonwood (*Populus deltoides*) as a ‘phytometer’¹⁰ to integrate the net growth response to multiple anthropogenic environmental changes in New York City compared to surrounding rural environments. Rapid growth rates, continuous growth throughout the season, and responsiveness to a range of climatic and pollutant variables^{11–15} make this widespread riparian and early successional tree species a suitable indicator. Soil transplants, nutrient budgets, chamber replication of field conditions and multiple regression approaches were then used to determine the key driving variables. Urban and rural site comparisons were selected from known steep pollution gradients^{1,16} across relatively short spatial scales (~100 km). Local variation in light and precipitation was minimized by growing plants in open fields with drip irrigation. Temperature effects on season length were controlled by synchronizing transplant and

[‡] Present addresses: US EPA Western Ecology Division, Corvallis, Oregon 97333, USA (J.W.G.);

Department of Integrative Biology, University of California, Berkeley, California 94720, USA (T.E.D.).

Table 1 Urban and rural atmospheric pollutants near New York City.

	Urban	Rural
Atmospheric gases (p.p.b.): $\bar{x}$ annual mean ($\pm$ s.e.)		
SO ₂	18.7 (0.3)	2.3 (0.0)
NO	39.3 (3.5)	0.5 (0.06)
NO ₂	37.7 (0.7)	6.2 (0.25)
O ₃ *	16.0 (1.5)	28.0 (0.6)
CO ₂ †	408 (0.2)	358 (0.4)
Suspended particulates (>10 μ g, μ g m ⁻³): $\bar{x}$ annual mean ($\pm$ s.e.)		
Pb	0.09 (0.00)	0.04 (0.00)
NO ₃ ⁻	5.47 (0.4)	0.44 (0.04)
SO ₄ ²⁻	12.4 (0.8)	4.3 (0.2)
Total	57.3 (5.2)	19.4 (2.3)
Wet deposition (mg m ⁻²): $\bar{x}$ third quarter total ($\pm$ s.e.)		
SO ₄ ²⁻	913.8 (151.2)	725.6 (65.9)
NO ₃ ⁻	519.9 (74.4)	465.7 (31.6)
NH ₄ ⁺	142.2 (10.0)	60.7 (6.9)
PO ₄ ³⁻	0.6 (0.01)	0.8 (0.1)
K ⁺	2.5 (1.4)	2.8 (0.6)
Ca ²⁺	59.5 (31.9)	17.0 (4.0)
Mg ²⁺	14.8 (8.8)	5.1 (1.0)
Na ⁺	72.9 (36.3)	14.4 (2.1)
Cl ⁻	133.7 (17.3)	41.3 (4.4)
H ⁺	18.0 (4.9)	19.0 (1.1)
pH‡	4.3 (0.1)	4.2 (0.1)

Atmospheric pollutant concentrations at urban and rural sites over the three years of experiments. Urban atmospheric gases, suspended particulates and wet deposition data were monitored at the New York State Department of Environmental Conservation's Morrisania, Green Point and Eisenhower Park air monitoring stations¹⁶, respectively (adjacent to sites NY_{1,3+4}, locations in Methods and Fig. 3b). Rural data are from HV₁¹⁹, with the exception of SO₂, Pb and CO₂ which were available from Belleayre¹⁶ (~70 km northwest of HV₁), Wallkill¹⁶ (~70 km southwest of HV₁) and LI₁¹⁷. Bold values show exposures that were significantly higher (paired *t*-tests, *P* < 0.05). Italics indicate data available in only one year in which case statistics represent intra-annual comparisons. Rural NO_x and urban/rural CO₂ concentrations were collected in years subsequent to field experiments, but follow well-documented urban-rural patterns^{4,5}. *May–September; †14 h, p.p.m., August 1996¹⁷. ‡ Precipitation weighted mean.

harvest dates within each growing season. The remaining microclimatic and pollutant differences were monitored at adjacent climate and air quality monitoring stations (<http://climod.nrc.cornell.edu>, <http://www.dec.state.ny.us>, and <http://www.ecostudies.org>).

The relative importance of atmospheric (that is, pollutants and residual microclimatic variation) versus soil effects was determined by growing cottonwoods in soils reciprocally transplanted from remnant urban and rural primary forest stands previously shown to differ in pH, hydrophobicity, conductivity, heavy metals (including Pb, Ni, S, Zn, Cu, Al and Mn) and base cation concentrations (Ca²⁺ and Mg²⁺)^{2,17}. Potting soil with slow-release fertilizer was also used

to estimate maximum growth potential at all sites independent of soil transplants. Given the net growth responses to site and soil treatments, we used the nutrient budgets to assess effects of wet and dry deposition; the chamber experiments to simulate urban and rural thermal, CO₂ and O₃ environments; and multiple regression analyses to relate final season biomass from field experiments to the remaining variables potentially responsible for observed growth differences.

Contrary to expectations, cottonwoods grew twice as large amid the high concentration of multiple pollutants in New York City compared to rural sites (Fig. 1). Greater urban plant biomass was found for all urban–rural site comparisons, two separate planting dates in the first year and two further consecutive growing seasons. Urban–rural growth differences occurred for the faster-growing trees in the fertilized potting soils and the slower-growing trees in the forest soil treatments, but there was no significant effect of soils transplanted from urban versus rural forests (*F*_{1,263} = 2.4, *P* = 0.124). The consistently greater urban plant biomass, independent of soil type, indicated that growth differences between urban and rural sites were due to atmospheric rather than soil alterations.

The beneficial effects of increased nutrient deposition or higher urban temperatures and CO₂ concentrations were primary factors potentially responsible for an increase in plant growth in the urban atmosphere (Table 1). However, the twofold growth differences between urban and rural sites occurred for the fertilized potting soil treatments where cottonwoods had access to three to five orders of magnitude more N, P, K⁺, Ca²⁺ and Mg²⁺ in fertilizer than from atmospheric deposition (11.1, 4.8, 9.2, 3.2 and 0.2 g respectively in fertilizer, compared to 28.1, 0.07, 0.31, 7.3 and 1.8 mg deposited to urban plants, respectively). Furthermore, atmospheric nutrient inputs were highest in year 2 when saplings grew the least (for example 2.2, 2.7 and 2.0 kg N ha⁻¹ per season via wet deposition at study site NY₄ for years 1, 2 and 3 respectively), and PO₄³⁻ showed a trend toward higher deposition at the rural sites (Table 1). The fertilization effects of nutrient deposition therefore did not appear to account for greater urban cottonwood biomass.

While temperature effects on season length were controlled via simultaneous transplant and harvest dates, a greater increase in C gain relative to respiratory C losses at the warmer daily temperatures (+1.8°C mean growing period temperatures¹⁷) could have enhanced cottonwood growth in the urban environment¹⁸. Because the relationship between C fixation and CO₂ concentration

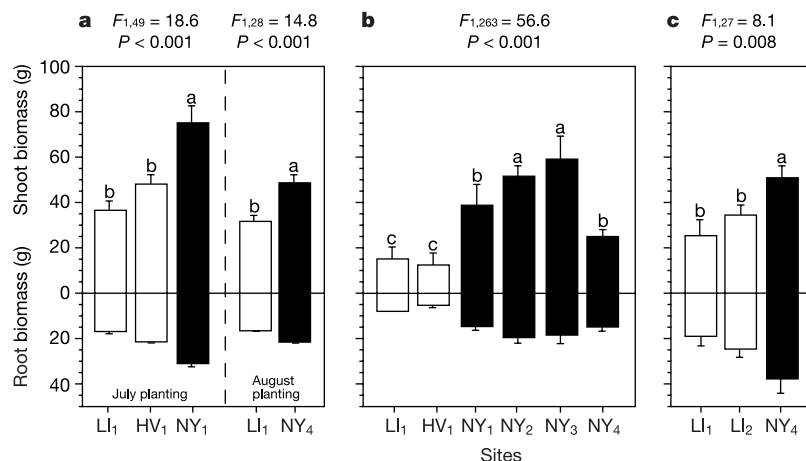


Figure 1 Cottonwood growth in urban and rural sites. Final season shoot and root biomass (mean $\pm$ s.e., potting soils) for cottonwoods grown in urban (filled, NY_{1–4}) and rural (open; HV₁, LI_{1–2}) sites in the vicinity of New York City for three consecutive growing seasons (**a–c**). Site locations in Methods and Fig. 3b. Values that fall below the zero line

are for belowground biomass. *F* and *P* statistics are for linear contrasts of analyses of variance comparing total biomass for urban versus rural sites. Independent comparisons for above- and belowground biomass gave the same result. Bars with different letters indicate values significantly different using the Tukey–Kramer HSD.

increases most with the initial rise in CO₂ above ambient conditions¹⁸, incrementally elevated urban CO₂ concentrations (+50 p.p.m.) could also have increased growth in urban sites. However, a series of chamber experiments simulating urban and rural thermal and CO₂ environments failed to reveal individual or combined effects of elevated urban temperatures and CO₂ concentrations on total ($F_{1,1-\text{Temp}} = 0.74$, $P = 0.55$; $F_{1,1-\text{CO}_2} = 0.67$, $P = 0.56$; $F_{1,1-\text{Temp} + \text{CO}_2} = 1.55$, $P = 0.43$), above- or below-ground biomass. The absence of a CO₂ effect is in agreement with other studies on this clone at higher CO₂ concentrations¹³. Multiple regression analysis also failed to reveal any relationship between final season biomass in urban and rural field sites and ambient temperature regimes whether the data were summarized as the maximum, minimum or mean of the daily temperature profiles or as cumulative growing degree-days (base 15 °C; $P = 0.882$, 0.895, 0.226 and 0.160, respectively; forward stepwise regression analysis).

The urban pollution haze can reduce maximum incoming radiation (1,800 $\mu\text{mol m}^{-2} \text{s}^{-1}$ photosynthetically active radiation) by up to 18%¹⁹, but light levels remained far above photosynthetic saturation for this species (700 $\mu\text{mol m}^{-2} \text{s}^{-1}$)¹². High concentrations of condensation nuclei can also increase precipitation in urban centres³, but irrigation waters were supplied at a rate far in excess of precipitation (6.2 cm H₂O d⁻¹) and there were no consistent differences in precipitation between our urban and rural sites ($F_{1,10} = 0.85$, $P = 0.36$). Whereas lower relative humidity³ (-10.3% mean growing period comparisons¹⁷), increased CO₂ concentrations and even the pollutants themselves could all have reduced stomatal conductance in the urban sites, thereby minimizing pollutant impacts¹⁸, the offset of detrimental impacts could not account for a relative increase in plant growth in urban compared to rural sites.

Overall, then, the collective results of all experiments and environmental comparisons provided no evidence that greater urban cottonwood biomass was due to enhanced growth in the urban atmosphere. Yet the same pattern could have arisen if detrimental effects reduced growth in the country. Because nutrient budgets, temperature and CO₂ experiments, temperature regressions and microclimatic comparisons could not account for

an increase in plant growth in the city, clearly these factors also could not account for reduced growth in the country. As expected, most atmospheric gases, suspended particulates and wet deposition components that could reduce plant growth were either higher in New York City or did not differ between urban and rural sites (Table 1). However, O₃ was significantly higher at rural sites, and thus could have reduced growth in the country. Primary O₃ precursors are emitted in cities, but must react in sunlight to form O₃ as air masses move to rural environments²⁰. Ozone exposures were therefore consistently higher for rural sites both to the north and the east of the city in all consecutive growing seasons (paired *t*-tests, $P < 0.001$).

An open-top chamber experiment exposing cottonwood to ambient and greater O₃ exposures representative of those at the urban and rural sites (33 versus 59 p.p.b. growing period mean

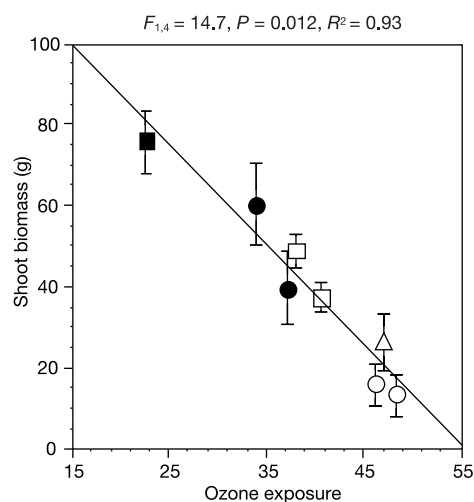


Figure 2 Cottonwood biomass related to O₃ exposure. Final season cottonwood shoot biomass (mean $\pm$ s.e., potting soils) at urban (filled) and rural (open) field sites versus ambient O₃ exposure (growing period 12-hour mean, p.p.b.; data available for NY₁, NY₃, HV₁ and LI₁; ref. 16). Squares, circles and triangles represent years 1, 2 and 3, respectively. *F* and *P* statistics show significance of O₃ effects from a multiple regression analysis with growing degree-days (base 15 °C; $F_{1,4} = 0.2$, $P = 0.685$) and urban versus rural sites ($F_{1,4} = 0.04$, $P = 0.866$) as additional independent variables. Variance inflation factors <10 demonstrated the lack of collinearity.

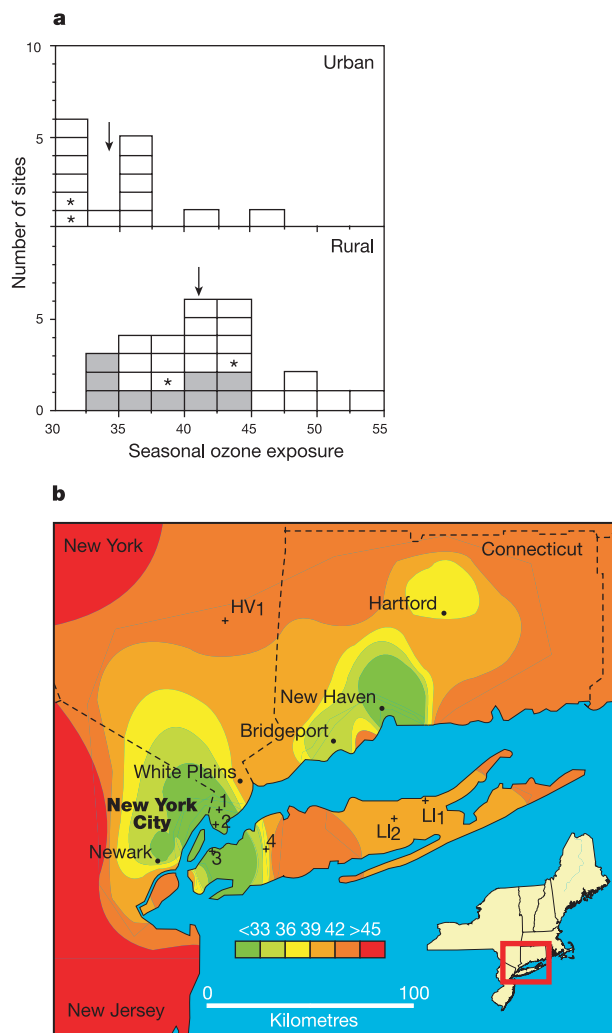


Figure 3 Urban and rural O₃ exposures in the northeastern USA. **a**, Histograms of O₃ exposures (12-hour mean p.p.b., May–September, year 2) for all urban and rural sites in the US EPA AIRS database for the northeastern USA (excluding Maine and Pennsylvania; rural grey: forested; rural white: agricultural). Exposures were significantly lower in urban compared to rural sites ($t_{41} = 4.1$, $P < 0.001$). Arrows show mean exposures. Asterisks show exposures at sites used in this study. Results were consistent for mean and cumulative peak O₃ (AOT40 or SUM06⁷) comparisons. **b**, Inverse distance weighted interpolation of O₃ exposure (units as above) for all US EPA AIRS sites in the New York/New Jersey/Connecticut region in year 2. The lower cumulative urban exposures appear as the green-yellow ‘footprints’ up to 500 km² in size. Note that footprints of lower urban O₃ appear only for areas with extensive urban, suburban and rural monitoring stations. Urban (NY_{1–4}), Hudson Valley (HV₁) and Long Island (LI_{1–2}) site locations are also shown.

concentrations, respectively) showed a 50% reduction in cottonwood biomass at the greater O₃ exposures ($F_{1,7} = 10.0$, $P = 0.016$)—an effect magnitude comparable to that found between urban and rural field sites. Multiple regression analysis showed that final season biomass was significantly inversely related to ambient O₃ exposures across all field sites and years of experiments, accounting for 93% of variation (Fig. 2). Within-season comparisons also showed that incremental changes in leaf area production were inversely related to cumulative O₃ exposures ($F_{1,289 \text{ Yr1}} = 8.2$, $P = 0.004$; $F_{1,188 \text{ Yr2}} = 11.5$, $P < 0.001$) independent of effects due to site, soil, time through the season or growing degree-days. Less leaf area was produced in the sites with the highest O₃ exposures and this pattern held, despite the switch in sites with the highest exposures between measurement intervals.

Analysis of the US Environmental Protection Agency's Alliance of Information and Referral Systems (AIRS) database (<http://www.airs.org>) showed that O₃ exposures at our rural sites were representative of mean non-urban agricultural and forested exposures throughout the northeastern USA (Fig. 3a). Consequently, the detrimental effect of higher rural than urban O₃ exposures was not due to extreme high exposures downwind of an urban centre²⁰. Instead, urban exposures were significantly lower than non-urban sites throughout the region. Spatial interpolation of the O₃ data for the New York/New Jersey/Connecticut region revealed footprints consisting of relatively low cumulative O₃ exposures in urban areas with a background of higher regional exposures (Fig. 3b). While NO_x titration reactions have been shown to reduce O₃ within the urban core⁷, regional interpolations specifically omit the urban data²¹, so the footprints of lower cumulative urban O₃ exposures have not previously been documented.

Of the many factors that could affect plant growth in urban environments, soil factors, nutrient deposition, temperature, CO₂, urban air pollutants and microclimatic variables could not account for the greater urban plant biomass. Rather, all evidence indicated that the greatest effect of the multiple anthropogenic environmental changes was the secondary reactions: these produced the higher rural O₃ exposures that reduced growth in the country. Because cottonwood is mid-range in O₃ sensitivity, with many species showing greater responses to ambient O₃ exposures^{22–24}, reduced growth in response to higher rural O₃ exposures is unlikely to be restricted to this cottonwood clone. These results do not negate the known detrimental effects of multiple urban pollutants, but show the greater effect of secondary reactions that create higher cumulative O₃ exposures beyond the urban core. Although individual 1-hour peak concentrations are typically higher in urban centres⁷, our data indicate that the higher cumulative exposures at rural sites had the greatest impact.

Our research thus determines the relative importance of multiple anthropogenic environmental changes under current field conditions, and reveals a number of counter-intuitive results. (1) There was greater plant growth amid multiple pollutants in urban compared to rural environments. (2) Higher urban temperatures, CO₂ concentrations and N deposition could not account for increased growth in the city. (3) Ozone was the single overriding factor accounting for observed growth differences among multiple anthropogenic environmental changes. (4) The most detrimental effects of multiple urban pollutant emissions occurred in rural environments, with (5) footprints of reduced impact of lower cumulative urban O₃ exposures on a background of higher regional exposures.

These findings are in contrast to the extensive monitoring, warning and effects research within city centres, suggestions that ecological differences between urban and rural environments could be due to higher urban than rural O₃ exposures²⁵, and the pervasive perception that rural environments are safe havens from urban pollutant emissions. As such, our work highlights the need to reconsider relative pollutant impacts in urban and rural environ-

ments as air sheds merge throughout the globe. Although extensive global change research has studied potential impacts of temperature, CO₂ and N deposition, this study shows overriding O₃ impacts amid these other factors. □

Methods

Field comparisons

Cottonwood growth in New York City was compared to a northern rural site in the Hudson Valley (HV) and eastern rural sites on Long Island (LI) for three consecutive growing seasons (1992–1994). Rooted cuttings (Clone ST109; initial height 10–25 cm; 5–12 leaves) were transplanted to the field on 4–7 July, drip irrigated (3.8 litres day⁻¹ in four intervals at 06:00, 10:00, 14:00 and 18:00 hours) and harvested before bud set and leaf senescence on 13–15 September. A second planting was also made on 6 August in year 1. Final season biomass comparisons were supplemented with measurements of total leaf area ($\text{area} = 3.772 - 1.611 \times \text{length} + 0.745 \times \text{length}^2$, $r^2 = 0.976$, $P < 0.001$) measured biweekly on all plants in year 1 and tri-weekly on three plants per soil treatment in year 2. Urban and rural sites (locations shown in Fig. 3b) were located at The New York Botanical Garden, Bronx (NY₁); Hunts Point Water Works, Bronx (NY₂); Con Edison Fuel Depot, Astoria (NY₃); Eisenhower Park, Hempstead (NY₄); The Institute of Ecosystem Studies, Millbrook (HV₁); Cornell University Horticultural Research Laboratory, Riverhead (LI₁); and Brookhaven National Laboratory, Upton (LI₂). Herbivory was negligible (<1% total leaf area) and did not vary between urban/rural site and soil treatments (ANOVA, $P > 0.5$).

Soils were collected from the organic and top soil horizons of remnant oak-dominated urban and rural forest stands and transplanted to urban and rural sites. Soils were transplanted from one urban and one rural forest to all sites in year 1 (five plants per soil origin), and from two urban and two rural forests from each of the HV and LI comparisons to all respective HV and LI sites in year 2 (four forest soils per site, ten plants per soil origin, eight forest soils in total). All soils were fine sandy loam: Charlton and Hollis series for the urban–rural HV comparisons and Montauk series for LI comparisons. Soils were sieved (1-cm mesh), mixed thoroughly, placed in 19-litre pots, transported to each site, and buried on 1-m centres. Holes were dug 50% below pot depth and backfilled with gravel and sand for drainage. HV soils were collected from Van Cortland and Pelham Bay Parks, Bronx, and Housatonic and Mohawk State Forests, northwestern Connecticut. LI soils were collected from Cunningham and Alley Pond Parks, Queens, The David Weld Preserve, Nissequogue, and Edward Stevenson's woods, Mt Sinai. The potting soil treatment, also used at all sites and all years of experiments (ten plants per site; 15 plants per site for the second planting of year 1), was perlite:topsoil:peat (1:2:1 v/v), with limestone (10.2 g per pot), 5N-10P-5K fertilizer (13.6 g per pot), phosphate (20%, 10.2 g per pot) and slow-release fertilizer (Osmocote 14N-14P-14K, Scotts-Sierra Co.; 113.4 g per pot). Site, soil and site $\times$ soil were the main effects in analyses for the first planting of year 1 and these factors were nested within HV and LI comparisons in year 2. Collinearity between time and growing degree-days for intra-annual foliar comparisons forced removal of the non-significant temperature effect from the regression model ($F_{1,286 \text{ Yr1}} = 0.00$, $P = 0.988$; $F_{1,187 \text{ Yr2}} = 1.4$, $P = 0.232$).

Nutrient budgets

N, P, K⁺ and base-cation deposition were calculated from NH₄⁺-N, NO₃⁻-N, PO₄³⁻-P, K⁺, Ca²⁺ and Mg²⁺ concentrations in precipitation¹⁶ with dry deposition assumed to equal that in rainfall²⁶. Nutrient comparisons in otherwise clear and odourless irrigation waters showed that As³⁺⁵⁺, Ba²⁺, Cd²⁺, Cr²⁺³⁺, Cu²⁺, F⁻, Pb²⁺³⁺, 4+, Se²⁺²⁺, Ag⁺ and Zn²⁺ were all below detection limits. The remaining detectable constituents (NH₃, NO₃⁻, Ca²⁺, Fe²⁺³⁺, K⁺, Mg²⁺, Mn²⁺³⁺, Na⁺, SO₄²⁻, Cl⁻, alkalinity, hardness, total solids, pH, anion-cation balance, conductivity and turbidity) were not different between urban and rural sites (t -tests, $P > 0.05$), and were not related with phytometer biomass (forward stepwise regression analysis, $P > 0.25$).

Chamber experiments

Field conditions were simulated as closely as possible by transplanting cottonwoods into the same 19-litre pots with fertilized potting soils and providing full sun and an ample water supply (biweekly irrigation to field capacity for open-top chambers and the same drip irrigation for chamber experiments). The open-top chamber O₃ experiment was performed at the Boyce Thompson Institute Field Facility in Ithaca, New York, from 7 July to 21 September 1996, in conjunction with an experiment that included four chambers per treatment but showed no significant chamber effects ($N = 5$ plants per chamber)²⁷. Temperature and CO₂ growth chamber experiments (Conviron PGW36, Controlled Environments, Inc.) simulated the temperatures and CO₂ concentrations of the NY₁ and LI₁ sites in year 2. Since cottonwoods showed twofold growth differences between sites in the first three weeks of the field experiments¹⁷, chamber conditions were set to the average conditions of the first 21 days and experiments lasted three weeks. Control conditions were: min. 18.8 °C/ max. 30.3 °C temperatures (ramped linearly between 6:00 and 15:00), and 350 $\pm$ 3 p.p.m. CO₂ concentrations. Treatments were asymmetrically elevated temperatures (min. 21.1 °C/ max. 31.9 °C) and elevated CO₂ concentration (400 $\pm$ 3 p.p.m.). Days were set to 14.5 hours, relative humidity was 50%, and light was maintained above 900 $\mu\text{mol m}^{-2} \text{s}^{-1}$ with dawn and dusk simulated by ramping chamber lights in 25% increments at 15-min intervals. Each experiment was repeated twice, switching the treatments between chambers ($N = 7$ plants per chamber) with treatment effects tested against the treatment by replicate interaction²⁸. Further analyses using individual plants as the experimental unit confirmed the absence of temperature and CO₂ effects.

Received 21 February; accepted 28 April 2003; doi:10.1038/nature01728.

1. Lovett, G. M. *et al.* Atmospheric deposition to oak forests along an urban-rural gradient. *Environ. Sci. Technol.* **34**, 4294–4300 (2000).
2. Pouyat, R. V., McDonnell, M. J. & Pickett, S. T. A. Soil characteristics of oak stands along an urban-rural land-use gradient. *J. Environ. Qual.* **24**, 516–526 (1995).
3. Peterson, J. T. *The Climate Of Cities: A Survey Of The Recent Literature*. NAPCA Pub. No. AP-59 (US Department of Health, Education and Welfare, 1969).
4. Idso, C. D., Idso, S. B. & Balling, R. C. Jr An intensive two-week study of an urban CO₂ dome in Phoenix, Arizona, USA. *Atmos. Environ.* **35**, 995–1000 (2001).
5. Gatz, D. F. Urban precipitation chemistry: A review and synthesis. *Atmos. Environ. B Urban Atmos.* **25**, 1–16 (1991).
6. Freedman, B. *Environmental Ecology: The Ecological Effects Of Pollution, Disturbance, And Other Stresses* (Academic, San Diego, 1995).
7. US National Research Council. Committee on Tropospheric Ozone Formation and Measurement. *Rethinking The Ozone Problem In Urban And Regional Air Pollution* (National Academy Press, Washington DC, 1992).
8. Pickett, S. T. A. *et al.* Urban ecological systems: Linking terrestrial ecology, physical, and socioeconomic components of metropolitan areas. *Annu. Rev. Ecol. Syst.* **32**, 127–157 (2001).
9. Mooney, H. A., Winner, W. E., Pell, E. J. & Chu, E. (eds) *Response Of Plants To Multiple Stresses* (Academic, San Diego, 1991).
10. Weinstein, L. H., Laurence, J. A., Mandl, R. H. & Walti, K. *Use of Native And Cultivated Plants As Bioindicators Of Pollution Damage: Plants For Toxic Assessment* (eds Wang, W., Gorsuch, J. W. & Lower, W. R.) 117–126 (Am. Soc. Testing and Materials, Philadelphia, 1990).
11. USDA Forest Service. *Cottonwood: Populus deltoides: hybrid poplar*. (Govt document no. A 13.42/24:C 82, Government Printing Office, Washington DC (1980)).
12. Wait, D. A., Jones, C. G. & Schaedle, M. Controlling growth and chemical composition of *Populus deltoides* saplings by iteratively matching nutrient supply to demand: a bootstrap fertilization technique. *Tree Phys.* **16**, 359–365 (1996).
13. Wait, D. A., Jones, C. G., Wynn, J. & Woodward, F. I. The fraction of expanding to expanded leaves determines the biomass response of *Populus deltoides* to elevated CO₂. *Oecologia* **121**, 193–200 (1999).
14. Coleman, J. S. & Jones, C. G. Plant stress and insect performance: cottonwood, ozone and a leaf beetle. *Oecologia* **76**, 57–61 (1988).
15. Magnussen, S. Growth, transpiration, unit leaf area, and water use efficiency of six poplar clones at different light and temperature regimes. *Forest Tree Improvement* **0**, 1–47 (1985).
16. New York State Department of Environmental Conservation, Division of Air Resources. *NY State Air Quality Report: Ambient Air Monitoring System*. Report no. DAR-95-1 (NYS-DEC, Albany, New York, 1995).
17. Gregg, J. W. *The Effect Of Urban Compared To Rural Environments On Plant Growth In The Vicinity Of New York City: The Net Response And Relative Importance Of The Different Pollutants*. Thesis, Cornell Univ (1999).
18. Jones, H. G. *Plants and microclimate: A quantitative approach to environmental plant physiology*. (Cambridge Univ. Press, 1996).
19. Jacovides, C. P., Timbíos, F., Asimakopoulou, D. N. & Steven, M. D. Urban aerosol and clear sky spectra for global and diffuse photosynthetically active radiation. *Agric. Forest Meteorol.* **87**, 91–104 (1997).
20. Hestvedt, E., Isaksen, I. S. A. & Hov, O. Ozone generation over rural areas. *Environ. Sci. Technol.* **12**, 1279–1284 (1978).
21. Lefohn, A. S., Knudsen, H. P., Logan, J. A., Simpson, J. & Bhumralkar, C. An evaluation of the kriging method to predict 7-h seasonal mean ozone concentrations for estimating crop loss. *J. Air Pollut. Control Assoc.* **37**, 595–602 (1987).
22. USDA Forest Service FIA Ozone Biomonitoring Program. *Field Methods For Phase 3 Measurements*. (USDA Forest Service, Amherst, Massachusetts, 2002).
23. Federal Land Managers' Air Quality Related Values Workgroup (FLAG). *Phase I Report*. (US Forest Service, National Park Service, and Fish and Wildlife Service, Fort Collins and Boulder, Colorado, 2000).
24. Appleton, B. & Koci, J. *Trees For Problem Landscape Sites – Air Pollution*. Virginia cooperative extension Pub. no. 430-022 (Virginia Cooperative Extension, Blacksburg, Virginia, 2000).
25. Goldman, M. B., Groffman, P. M., Pouyat, R. V., McDonnell, M. J. & Pickett, S. T. A. CH₄ uptake and N availability in forest soils along an urban-rural gradient. *Soil Biol. Biogeochem.* **27**, 281–286 (1995).
26. Lovett, G. M. & Lindberg, S. E. Atmospheric deposition and canopy interactions of nitrogen in forests. *Can. J. Forest Res.* **23**, 1603–1616 (1993).
27. Yun, S. C. & Laurence, J. A. The response of clones of *Populus tremuloides* differing in sensitivity to ozone in the field. *New Phytol.* **141**, 411–421 (1999).
28. Potvin, C. & Tardif, S. Sources of variability and experimental designs in growth chambers. *Funct. Ecol.* **2**, 123–130 (1988).
29. Institute of Ecosystem Studies. *Environmental Monitoring Program* (Institute of Ecosystem Studies, Millbrook, NY, 2002).

Acknowledgements M. McDonnell, R. Pouyat, S. Pickett, A. Greller, G. Lovett, M. Geber and P. Marks provided discussions at the outset of this research. C. Andersen, J. Compton, A. Hudak, J. Laurence, H. Lee, D. Phillips, P. Rygielwicz, A. Solomon and D. Tingey provided discussions and editorial feedback. H. Lee provided EPA O₃ data and statistical consultation. P. Dickerson created the inverse-distance weighted O₃ map. M. Topa oversaw the open-top chamber experiment. Organizations listed in the Methods provided site access, forest soils and technical assistance. Financial support was provided to J.W.G. by the Edna Bailey Sussman Fund for Environmental Internships, the New York State Heritage Foundation, a Cornell University Mellon Research Grant, the Institute of Ecosystem Studies, Cornell's Department of Ecology and Systematics, a Mellon Foundation graduate training grant (to T.E.D.), the Cornell Center for the Environment, Sigma Xi and a US EPA post doctoral fellowship.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.W.G. (jwg1@cornell.edu).

Strong population substructure is correlated with morphology and ecology in a migratory bat

Cassandra M. Miller-Butterworth^{*†‡}, David S. Jacobs^{*} & Eric H. Harley[†]

^{*} Department of Zoology, University of Cape Town, Private Bag, Rondebosch, 7701, South Africa

[†] Wildlife Genetics Unit, Division of Chemical Pathology, University of Cape Town, Observatory, 7925, South Africa

Examining patterns of inter-population genetic diversity can provide valuable information about both historical and current evolutionary processes affecting a species. Population genetic studies of flying and migratory species such as bats and birds have traditionally shown minimal population substructure, characterized by high levels of gene flow between populations^{1,2}. In general, strongly substructured mammalian populations either are separated by non-traversable barriers or belong to terrestrial species with low dispersal abilities³. Species with female philopatry (the tendency to remain in or consistently return to the natal territory) might show strong substructure when examined with maternally inherited mitochondrial DNA, but this substructure generally disappears when biparentally inherited markers are used, owing to male-mediated gene flow⁴. Male-biased dispersal is considered typical for mammals⁵, and philopatry in both sexes is rare. Here we show strong population substructure in a migratory bat species, and philopatry in both sexes, as indicated by concordance of nuclear and mtDNA findings. Furthermore, the genetic structure correlates with local biomes and differentiation in wing morphology. There is therefore a close correlation of genetic and morphological differentiation in sympatric subspecific populations of this mammalian species.

Schreibers' long-fingered bat, *Miniopterus schreibersii natalensis* (Chiroptera, Vespertilionidae), migrates seasonally between wintering roosts (hibernacula) and summer maternity colonies in South Africa. Ringing studies⁶ indicate that fidelity to both roost types is well developed. *Miniopterus schreibersii natalensis* were sampled (Fig. 1, Supplementary Table S1) from four South African maternity roosts (Die Hel (DHL), Jozini Dam (JD), Peppercorn (PC) and Sudwala (SW)), one hibernaculum (Steenkampsdraai (SKK)), four roosts that are occupied all year round but are used primarily as summer roosts (De Hoop (DHP), Koegelbeen (KB), Grahamstown (G) and Maitland Mines (MM)) and one transient pre-maternity roost (Shongweni Dam (SHD)). An analysis of six dinucleotide microsatellite loci indicated that the *M. s. natalensis* population was genetically substructured into three major subpopulations, occurring in the south (DHL and DHP), west (SKK and KB) and northeast (G, MM, SHD, JD, PC and SW) regions of the country (Fig. 2, Supplementary Table S2). Colonies within each subpopulation were genetically similar and thus poorly differentiated: ρ values ranged between -0.005 and 0.068 ($P > 0.05$ for all comparisons). Genetic distances between these colonies were also low: $(\delta\mu)^2$ ranged between 0.084 and 0.446 . However, colonies from different subpopulations were strongly differentiated, both when examined individually through pairwise comparisons (range of ρ values 0.152 – 0.686 , $P < 0.01$ for all comparisons; range of $(\delta\mu)^2$ values 1.033 – 5.286) and when pooled into the three subpopulations (range of ρ values 0.351 – 0.623 , $P \leq 0.0001$ for all comparisons; range of $(\delta\mu)^2$ values 2.037 – 4.550). Each colony had sufficiently

[‡] Present address: Laboratory of Genomic Diversity, National Cancer Institute, PO Box B, Frederick, Maryland 21702, USA.

characteristic genotype frequencies for an assignment test⁷ to be able to assign a significant majority of individuals correctly to both the colony (62–90%) and the subpopulation (95–100%) from which they were sampled (Supplementary Table S3, Supplementary Fig. S1).

The pattern of genetic differentiation obtained from mtDNA control region sequences was closely correlated with the microsatellite results. Fifty-five unique haplotypes were identified from 78 samples that were sequenced. Mismatch distribution and linkage disequilibrium analyses suggested that the South African *M. s. natalensis* population and the three subpopulations have either been stable or expanding during their recent history (Supplementary Information). Analysis of molecular variance indicated strong substructure within the *M. s. natalensis* population ($\Phi_{ST} = 0.620$, $P = 0.002$), with 56% of the variation occurring among subpopulations and less than 6% among colonies within subpopulations. The mean divergence of control region haplotypes among *M. s. natalensis* subpopulations was 3%, and some haplotypes differed by up to 6%. These values are within the upper range of those reported in very mobile vertebrates, such as marine turtles⁸ and ghost bats⁹. A sequence divergence of more than 3% is unusual in highly mobile species, which usually show less than 1% divergence, unless they are geographically isolated¹. Other flying, migratory species such as the Mexican free-tailed bat, *Tadarida brasiliensis mexicana*¹, and red-winged blackbirds, *Agelaius phoeniceus*², show genetic homogeneity and high levels of gene flow. Conversely, *M. s. natalensis* haplotypes are substructured into three geographically and genetically distinct assemblages (Fig. 3), which, with the exception of four (possibly ancestral) haplotypes that cluster outside their regional groups, are separated by minimum evolutionary divergences twice that of the most divergent haplotypes within subpopulations. This supports similar findings for *M. schreibersii* in Australia¹⁰.

In most mammals studied, including bats^{4,11–14}, males disperse and females are philopatric. However, the close agreement here between nuclear and mitochondrial markers suggests that both sexes of *M. s. natalensis* show strong philopatry to their natal subpopulations, leading to restricted gene flow among the subpopulations. In contrast, the low levels of genetic differentiation within subpopulations suggest extensive gene flow between these colonies. This species might therefore display similar behaviour to that suggested for pilot whales, *Globicephala melas*¹⁵, and the brown

long-eared bat, *Plecotus auritus*¹⁶. In these species, both sexes remain within their natal pod or colony but mate with individuals from other groups. In *M. s. natalensis*, mating is believed to occur at the hibernacula⁶, and it is possible that individuals from several maternity colonies within a subpopulation use a common hibernaculum, as has been observed in Portugal¹⁷. This would facilitate gene flow between colonies within a subpopulation, while limiting gene flow to other subpopulations.

There seem to be no zoogeographic barriers to *M. s. natalensis* dispersal across subpopulations in South Africa. Furthermore, Mantel tests indicated that neither genetic differentiation, ρ , nor genetic distance, $(\delta\mu)^2$, was significantly correlated with geographic distances between the colonies, when examined across the entire country ($r = 0.254$, $P = 0.07$; and $r = 0.004$, $P = 0.44$, respectively), or within the largest (northeastern) subpopulation ($r = 0.439$, $P = 0.07$; and $r = 0.179$, $P = 0.202$, respectively). A significant correlation was found between mtDNA haplotype differentiation (Φ_{ST}) and geographic distance between colonies, both within South Africa as a whole ($r = 0.474$, $P = 0.005$) and within the northeastern subpopulation ($r = 0.730$, $P < 0.001$). This indicates that female-mediated gene flow might be explained partly by isolation by distance, particularly within the largest subpopulation. However, the mtDNA correlation across the entire country is weak, and isolation by distance explains only a small proportion ($r^2 = 0.225$) of the genetic variation. Migration or dispersal between SKK and DHL, for example (~170 km), is well within the ability of these bats, but these colonies are genetically very distinct ($\rho = 0.498$; $(\delta\mu)^2 = 3.419$) and movement between them is limited. Clearly factors other than physical distance function in determining how individuals move between colonies.

Although morphological characters are influenced by selection and therefore evolve more slowly, providing less resolution than neutral genetic data, the variation in wing morphology between the different *M. s. natalensis* colonies broadly supports the pattern of population differentiation indicated by the genetic markers. Principal components analysis (PCA) and analysis of variance (ANOVA) of the morphological data yielded two groups of colonies (Fig. 4). Colonies in the northeastern group (SHD, JD, PC, SW and KB), which had high factor scores on principal component 1, had significantly higher aspect ratios (ARs) than those of the southwestern group (DHP, DHL, G, MM and SKK), which had low factor scores on principal component 1. KB had intermediate ARs; it differed significantly in AR from DHP, DHL, MM and G, but not

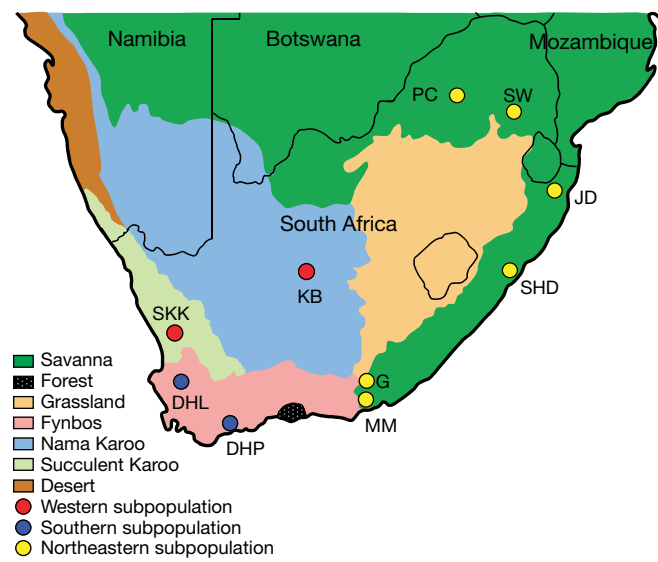


Figure 1 Location of major *M. s. natalensis* colonies and subpopulations in relation to the major biomes²⁰ of South Africa. Scale bar, 200 km.

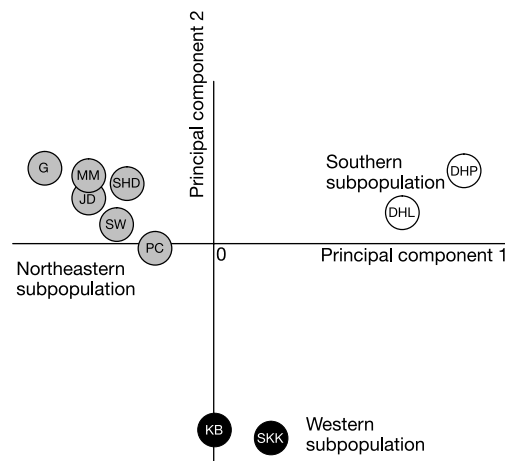


Figure 2 PCA of *M. s. natalensis* colony differentiation, generated from microsatellite data in PCA-GEN. The first two principal components (PC1 and PC2) plotted here account for 85% of the variance (inertia) in the data (per-axis inertia: PC1 = 0.353; PC2 = 0.211).

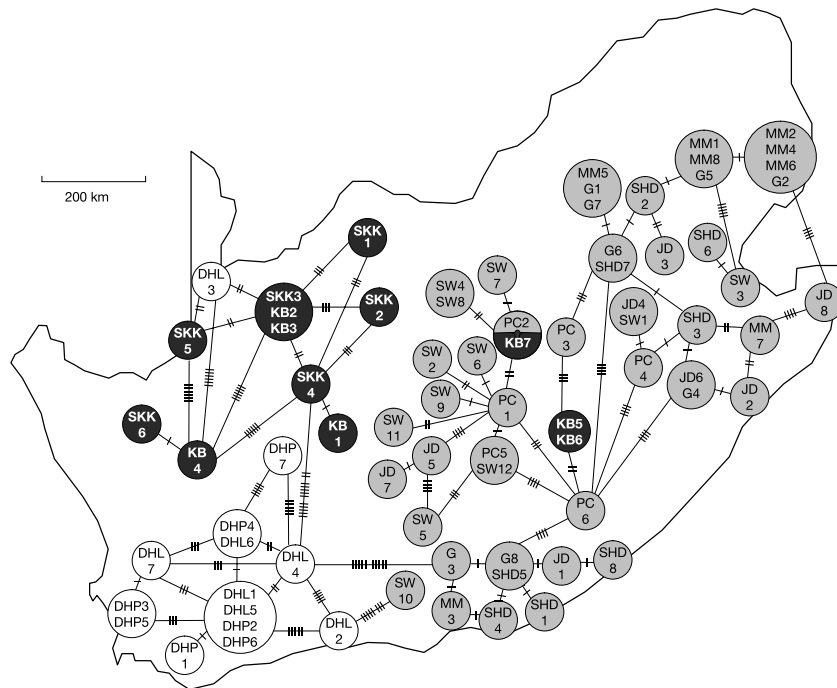


Figure 3 Minimum spanning network of 55 unique mtDNA control region sequence haplotypes identified in *M. s. natalensis* colonies, superimposed over a map of South Africa. Haplotypes are clustered into three regional subpopulations: southern, western and northeastern. Haplotype names indicate the colony where the individual was captured and the unique number of the haplotype within the colony. Haplotypes are shaded

according to the subpopulation in which the individuals were captured: white, southern; black, western; grey, northeastern. The extent of evolutionary divergence (connection length) of haplotypes is indicated by the number of cross-hatches on each line; lengths of the connection lines are not directly proportional to divergence distances.

from SKK or any of the northeastern colonies (Supplementary Table S4).

A possible explanation for these morphological differences relates to the migratory characteristics of this species. The patterns of genetic variation that we found, as well as previous behavioural studies⁶, indicate that bats in the north of the country migrate seasonally over hundreds of kilometres. Some bats from KB migrate to SKK to hibernate, a distance of ~560 km, considerably farther than has previously been recorded for this species in South Africa. Wings with a high AR experience lower drag during flight and improved flight efficiency over long distances¹⁸. The high ARs of the northern and KB bats might therefore reflect an adaptation to long-distance migration. In contrast, bats from DHL and DHP, and a large proportion of those from G and MM, migrate much shorter distances each year. A high AR might not be as great a benefit to them, particularly because it reduces manoeuvrability¹⁹.

The genetic and morphological differentiations seem to be associated with environmental conditions. The locations of the three genetic subpopulations correspond to four of the seven major biomes in South Africa²⁰, namely savanna, fynbos, Nama Karoo and succulent Karoo (Fig. 1), all of which were established during the Neogene period (25–2 Myr ago). The southern colonies, DHP and DHL, are located in the fynbos biome, an evergreen, sclerophyllous shrubland, with primarily winter rainfall. This biome replaced subtropical forests and became established in the early Pliocene (20–10 Myr ago)²⁰. The western colonies, SKK and KB, are found, respectively, in the succulent Karoo and Nama Karoo biomes. These are semi-arid, dwarf shrublands, which differ primarily in their relative proportions of succulent and grassy vegetation types²⁰. The colonies that make up the northeastern subpopulation (SHD, JD, PC, SW, G and MM) are found in the savanna, a primarily summer-rainfall area with vegetation types ranging from grasses and shrubs to woodlands²⁰. Whereas the relatively open and dry conditions of the Karoo biomes had already developed by the late Miocene

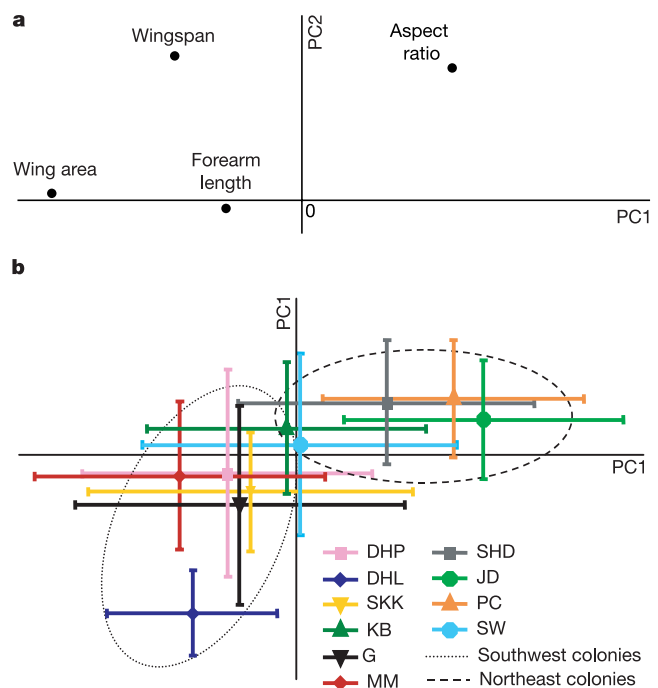


Figure 4 PCA of *M. s. natalensis* wing morphology. **a**, Factor loadings for wing morphology variables. Principal components PC1 and PC2 account for 76% of the variance in the data (eigenvalues: PC1 = 1.658; PC2 = 1.376). AR has a high factor score on PC1; all other variables have low factor scores on this axis. **b**, Mean component scores and standard errors for each colony. The overlap in error bars is expected for analysis of a single species for which complete morphological separation of colonies is unlikely. The colonies separate primarily along PC1, indicating that they differ in AR. ANOVA confirmed the existence of two geographical groups (southwest and northeast), which differ significantly in AR. This corresponds to the microsatellite and mtDNA findings, although KB groups morphologically with both SKK and the northeastern colonies.

(25–10 Myr ago), changes in the savanna biome are more recent and involved a change from dense woodland to more open grassland around Pliocene/Pleistocene times (10–1 Myr ago)²⁰. Thus, apart from cyclical vegetation changes during the Quaternary, which resulted in shifts in biome composition and boundaries²⁰, the modern biomes have been in existence for longer than a million years.

The close association between the locations of *M. s. natalensis* subpopulations and major biomes (Fig. 1) indicates that the bats might have adapted to local environmental conditions surrounding their roosts. This might involve adaptation to local vegetation types, to climatic conditions and/or to regional differences in prey abundance, which are closely correlated with rainfall²¹. The bats might respond to these regional differences by specializing on locally available prey, adapting their foraging strategies and/or reproductive cycles to ensure that parturition coincides with the time of the greatest availability of insects²². *Miniopterus schreibersii* is known to exhibit considerable plasticity in the timing of reproductive events, particularly the timing and duration of delayed implantation, in response to changing latitude, temperature, rainfall and food availability²³. Bats that are adapted to different biomes might therefore have sufficiently different reproductive cycles to restrict interbreeding between the subpopulations or biomes. Philopatry (in both sexes) to a subpopulation will be favoured because it enhances adaptation to the local environment, which might otherwise be lost through dispersal and associated gene flow.

Although the three subpopulations are not sufficiently distinct to merit separate species or even subspecies status, once gene flow becomes restricted, assortative mating and reproductive isolation can result, and speciation can ultimately occur. The *M. s. natalensis* subpopulations might therefore provide a mammalian example of incipient speciation involving mechanisms mediated by habitat and/or behaviour. □

Methods

Sample collection

Miniopterus schreibersii natalensis ($n = 307$) of both sexes (Supplementary Table S1) were sampled from 10 major colonies in South Africa between 1997 and 1999. Bats were captured either with a large hand-net while they were roosting during the day, or with a harp trap on emergence at night. Bats were placed temporarily in individual cloth bags and released as soon after processing as possible. Two skin biopsies 3 mm in diameter were taken²⁴ and stored in 500 μ l extraction buffer (0.5% SDS, 400 mM NaCl, 100 mM EDTA, 50 mM Tris-HCl pH 7.5) until extraction.

Genetics analysis

To determine microsatellite population parameters, each individual was examined at six dinucleotide microsatellite loci^{25,26} as described previously²⁵. Samples that did not amplify at four or more loci were excluded ($n = 6$). The number of alleles, the observed (H_O) and expected (H_E) heterozygosity per locus (Supplementary Information) and the frequency of each allele in each colony were determined with AGARst (E. H. Harley, University of Cape Town, South Africa). Each locus was tested for linkage disequilibrium and compliance with Hardy–Weinberg expectations by using GENEPOP²⁷. Critical significance levels were computed with sequential Bonferroni corrections.

For microsatellite population differentiation, a global estimate of ρ for *M. s. natalensis* in South Africa, pairwise ρ comparisons between colonies, and pairwise ϕ comparisons between sexes within individual colonies were performed, with the use of AGARst and R_{ST} Calc²⁸. ρ is a measure of population subdivision equivalent to R_{ST} but corrected for sample size differences. Permutation tests were implemented in R_{ST} Calc and critical significance values were computed with sequential Bonferroni corrections. The genetic distance, $(\delta\mu)^2$ between colonies was estimated with R_{ST} Calc. Pairwise ρ and $(\delta\mu)^2$ between colonies were also calculated separately for each sex and compared with Mann–Whitney rank sum tests. No significant difference between the sexes was found. PCA was performed with PCA-GEN (J. Goudet, University of Lausanne, Switzerland), incorporating 1,000 randomizations, to obtain a graphical representation of colony differentiation. Mantel tests, implemented in ARLEQUIN (S. Schneider, D. Roessli and L. Excoffier, University of Geneva, Switzerland), were used to determine whether $(\delta\mu)^2$ or ρ differences between colonies were correlated with geographic separation of colonies. An assignment test⁷ was performed in AGARst to indicate how characteristic any particular individual's multi-locus genotype was of the colony from which it was sampled (Supplementary Information).

For mitochondrial DNA population differentiation, ~530 base pairs of mitochondrial control region were sequenced in six to eight individuals from each location ($n = 78$), with published primers C and E¹³. Polymerase chain reactions (PCRs) were performed in an Eppendorf Mastercycler gradient in 50- μ l reactions, containing 0.03 U BIOTAQ DNA

polymerase (Bioline), 1.5 mM MgCl₂, 1 × BIOTAQ reaction buffer (16 mM ammonium sulphate, 67 mM Tris-HCl pH 8.8, 0.01% Tween 20), 0.2 mM dNTP mix (AP Biotech), 0.5 pmol primer and 20–50 ng template DNA. PCR cycles consisted of 2 min at 95 °C, 35 cycles of 50 s at 95 °C, 50 s at 50 °C and 1 min at 72 °C, and 10 min of extension at 72 °C. PCR products were purified with a QIAquick PCR purification kit (QIAGEN), cycle-sequenced with the Big Dye Terminator Cycle Sequencing Kit (Applied Biosystems) with primer P or E¹³, and subjected to electrophoresis through an ABI Prism 3100 Genetic DNA Analyzer (Applied Biosystems). The control region origin of the sequences was confirmed by a BLAST search.

Sequences were aligned manually with DAPSA (E. H. Harley, University of Cape Town, South Africa), and genetic distances between them were estimated with the Tamura–Nei plus gamma correction. Three hierarchical components of genetic diversity were examined through analysis of molecular variance: within colonies, among colonies within regions, and among regions. A minimum spanning network of haplotypes was constructed with ARLEQUIN. Mismatch distribution and linkage disequilibrium analyses were conducted to examine historical population demography (see Supplementary Information).

Morphological analysis

Wing tracings²⁹ and/or digital photographs (Olympus C-800L camera) of the extended right wing were obtained for each bat, in addition to age (adult or juvenile), sex, mass and forearm length. Wingspan and wing area were determined with BATWING³⁰. Aspect ratio and wing loading were calculated from these values¹⁸. Aspect ratios calculated from wing tracings were on average 7% greater than those determined from photographs, and were reduced accordingly so that these measurements could be compared directly.

PCA

PCA was performed on the variables wing area, wingspan, forearm length and aspect ratio of adults ($n = 286$) by using STATISTICA (StatSoft Inc.). Mass and wing loading were excluded because seasonal variation in sampling time caused differences in mass between colonies. Owing to the large number of bats sampled per colony, the mean component score for each colony, rather than individual values, was plotted in principal components morphological space. Variables identified by PCA as likely to differ between colonies were tested by one-way ANOVA, using MINITAB (Minitab Inc.). Dunn's test was performed to determine which colonies differed significantly from each other and which variables were responsible for those differences. Multifactor ANOVA was performed to confirm that no significant morphological differences existed between the sexes.

Received 4 February; accepted 15 April 2003; doi:10.1038/nature01742.

- McCracken, G. E., McCracken, M. K. & Vawter, A. T. Genetic structure in migratory populations of the bat *Tadarida brasiliensis mexicana*. *J. Mammal.* **75**, 500–510 (1994).
- Ball, R. M. J., Freeman, S., James, F. C., Bermingham, E. & Avise, J. C. Phylogeographic population structure of Red-winged blackbirds assessed by mitochondrial DNA. *Proc. Natl Acad. Sci. USA* **85**, 1558–1562 (1988).
- Avise, J. C. et al. Intraspecific phylogeography: The mtDNA bridge between population genetics and systematics. *Annu. Rev. Ecol. Syst.* **18**, 489–522 (1987).
- Petit, E. & Mayer, F. Male dispersal in the noctule bat (*Nyctalus noctula*): Where are the limits? *Proc. R. Soc. Lond. B* **266**, 1717–1722 (1999).
- Greenwood, P. J. Mating systems, philopatry and dispersal in birds and mammals. *Anim. Behav.* **28**, 1140–1162 (1980).
- van der Merwe, M. Aspects of the social behaviour of the Natal clinging bat, *Miniopterus schreibersii natalensis* (A. Smith, 1834). *Mammalia* **37**, 379–389 (1973).
- Paetkau, D., Waits, L. P., Clarkson, P. L., Craighead, L. & Strobeck, C. An empirical evaluation of genetic distance statistics using microsatellite data from bear (*Ursidae*) populations. *Genetics* **147**, 1943–1957 (1997).
- Norman, J. A., Moritz, C. & Limpus, C. J. Mitochondrial DNA control region polymorphisms: Genetic markers for ecological studies of marine turtles. *Mol. Ecol.* **3**, 363–373 (1994).
- Worthington Wilmer, J., Moritz, C., Hall, L. & Toop, J. Extreme population structuring in the threatened ghost bat, *Macroderma gigas*: Evidence from mitochondrial DNA. *Proc. R. Soc. Lond. B* **257**, 193–198 (1994).
- Cardinal, B. R. & Christidis, L. Mitochondrial DNA and morphology reveal three geographically distinct lineages of the large bentwing bat (*Miniopterus schreibersii*) in Australia. *Aust. J. Zool.* **48**, 1–19 (2000).
- Wilkinson, G. S. The social organization of the common vampire bat. II. Mating system, genetic structure and relatedness. *Behav. Ecol. Sociobiol.* **17**, 123–134 (1985).
- Kerth, G., Mayer, F. & König, B. Mitochondrial DNA (mtDNA) reveals that female Bechstein's bats live in closed societies. *Mol. Ecol.* **9**, 793–800 (2000).
- Wilkinson, G. S. & Chapman, A. M. Length and sequence variation in evening bat D-Loop mtDNA. *Genetics* **128**, 607–617 (1991).
- Thompson, M. J. A. Roost philopatry in female pipistrelle bats *Pipistrellus pipistrellus*. *J. Zool.* **228**, 673–679 (1992).
- Amos, B., Schlotterer, C. & Tautz, D. Social structure of pilot whales revealed by analytical DNA profiling. *Science* **260**, 670–672 (1993).
- Burland, T. M., Barratt, E. M., Beaumont, M. A. & Racey, P. A. Population genetic structure and gene flow in a gleaner bat, *Plecotus auritus*. *Proc. R. Soc. Lond. B* **266**, 975–980 (1999).
- Palmeirim, J. M. & Rodrigues, L. in *Ecology, Evolution and Behaviour of Bats* (eds Racey, P. A. & Swift, S. M.) 219–231 (Clarendon, Oxford, 1995).
- Norberg, U. M. & Rayner, J. M. V. Ecological morphology and flight in bats (Mammalia: Chiroptera): Wing adaptations, flight performance, foraging strategy and echolocation. *Phil. Trans. R. Soc. Lond. B* **316**, 335–427 (1987).
- Fenton, M. B. The foraging behaviour and ecology of animal-eating bats. *Can. J. Zool.* **68**, 411–422 (1990).

20. Cowling, R. M., Richardson, D. M. & Pierce, S. M. (eds) *Vegetation of Southern Africa* (Cambridge Univ. Press, 1997).
21. Rautenbach, I. L., Kemp, A. C. & Scholtz, C. H. Fluctuations in availability of arthropods correlated with microchiropteran and avian predator activities. *Koedoe* 31, 77–90 (1988).
22. Bernard, R. T. F. Reproductive synchrony and annual variation in foetal growth rate in the long-fingered bat (*Miniopterus schreibersii*). *J. Zool.* 232, 485–490 (1994).
23. Racey, P. A. & Entwistle, A. C. in *Reproductive Biology of Bats* (eds Crichton, E. G. & Kruttsch, P. H.) 363–414 (Academic, London, 2000).
24. Worthington Wilmer, J. & Barratt, E. M. A non-lethal method of tissue sampling for genetic studies of Chiropterans. *Bat Res. News* 37, 1–3 (1996).
25. Miller-Butterworth, C. M., Jacobs, D. S. & Harley, E. H. Isolation and characterisation of highly polymorphic microsatellite loci in Schreibers' long-fingered bat, *Miniopterus schreibersii* (Chiroptera: Vespertilionidae). *Mol. Ecol. Notes* 2, 139–141 (2002).
26. Moore, S. S., Hale, P. & Byrne, K. NCAM: a polymorphic microsatellite locus conserved across eutherian mammal species. *Anim. Genet.* 29, 33–36 (1998).
27. Raymond, M., Rousset, F. et al. GENEPOP (Version 1.2): Population genetics software for exact tests and ecumenicism. *J. Hered.* 86, 248 (1995).
28. Goodman, S. R. Calc: a collection of computer programs for calculating unbiased estimates of genetic differentiation and determining their significance for microsatellite data. *Mol. Ecol.* 6, 881–885 (1997).
29. Saunders, M. B. & Barclay, R. M. Ecomorphology of insectivorous bats: A test of predictions using two morphologically similar species. *Ecology* 73, 1335–1345 (1992).
30. Harley, E. H. & Miller-Butterworth, C. M. A software assistant for bat wing measurements. *Bat Res. News* 41, 99–102 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank R. Bernard, P. Taylor, B. Eick, C. Logan, M. Kirkman and R. Louw for assistance with obtaining samples; P. Owen, I. Baumgarten, C. van Heerden and D. James for technical assistance in the laboratory; and M. Butterworth and E. Teeling for constructive comments on the manuscript. This work was funded by grants from the National Research Foundation of South Africa and the University Research Committee of the University of Cape Town.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.M.B. (cbutterworth@mail.ncifcrf.gov). Microsatellite primer sequences and unique mitochondrial DNA control region haplotype sequences have been deposited in GenBank under accession numbers AY056588–AY056592 and AY228384–AY228439 respectively.

Captivity masks inbreeding effects on male mating success in butterflies

Mathieu Joron*† & Paul M. Brakefield*

* Institute of Biology, Leiden University, PO Box 9516, 2300 RA Leiden, The Netherlands

Small isolated populations are frequently genetically less diverse than core populations, resulting in higher homozygosity that can hamper their long-term survival^{1–4}. The decrease in fitness of organisms owing to matings between relatives is well known from captive and laboratory animals. Such inbreeding can have strongly deleterious effects on life-history traits and survival^{5–11}, and can be critical to the success of population conservation^{2,4,12}. Because pedigrees are hard to follow in the wild, most field studies have used marker loci to establish that fitness declines with increasing homozygosity^{1,13,14}. Very few have experimentally explored the effects of inbreeding in the wild¹⁵, or compared observations in the laboratory with field conditions^{8,9}. Here, using a technique involving the transfer of marker dusts during copulation, we show that a small decrease in mating success of captive inbred male butterflies in cages is greatly accentuated in conditions with unconstrained flight. Our results have important

implications for conservation and for studies of sexual selection because they show that the behaviours underlying patterns of mating can be profoundly influenced by a history of inbreeding or by any restraining experimental conditions.

Inbreeding depression in insects has been studied mainly in early life stages, especially with regard to egg-hatching and larval survival^{7,8}, and the issue of how behavioural characters are affected by inbreeding is largely unknown^{16,17}. Yet mating behaviour and courtship, with choosy females selecting for honest signalling by males, together with intense male–male competition, are likely to have central roles in variation in lifetime reproductive success^{9,14,18}, and could therefore represent an important part of the genetic load of populations in the wild.

Populations of the small African butterfly *Bicyclus anynana* have a high genetic load, especially for egg-hatching rates^{6,10}. The expression of inbreeding depression is important in determining the patterns of diversity in laboratory metapopulations¹⁰. Here we explore how the dynamics of inbreeding can directly influence pairings in a local population by performing mate-competition experiments, both as mating trials based on the choice of single females in the laboratory and at the population level in less constrained flight conditions in a large tropical greenhouse.

From the same outcrossed population, we first derived replicate groups of males of synchronized adult emergence with one of three inbreeding coefficients, $F = 0$ (F_0), $F = 0.25$ ($F_{0.25}$) and $F = 0.375$ ($F_{0.375}$). We then performed replicated mating trials in small cages in the laboratory, each involving competition between three males for an unmated, outbred female. Outbred (F_0) males showed a weak but repeatable advantage in mating success over either class of inbred male. This was significant when the inbred males were pooled ($G = 4.83$, d.f. = 1, $P = 0.027$; Fig. 1). Individual mating success in these constrained, artificial conditions was therefore dependent on the inbreeding history of the males. However, although we observed rejection behaviour by females, it is unclear whether they are able to assess the male's inbreeding coefficient as such. No difference in mating success was observed between the two levels of inbred males (Fig. 1).

Such mating trials enable the study of individual mating success, but the small volume of the cages is unlikely to allow the full expression of courtship characters. The courtship of *B. anynana*

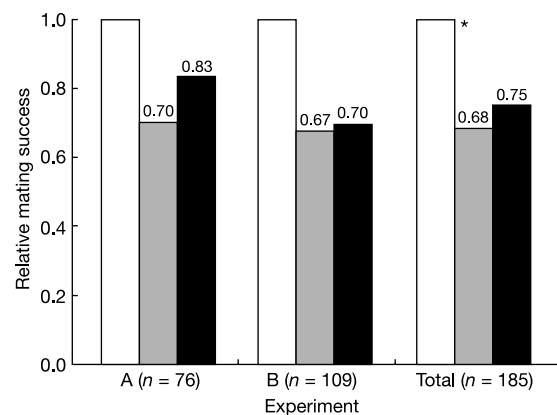


Figure 1 Mating success of inbred males relative to that of outbred males in laboratory trials. Two rounds of experiments, A and B, were performed with butterflies drawn at random from different sets of independent lines: seven F_0 lines (white bars), seven $F_{0.25}$ lines (grey bars) and six $F_{0.375}$ lines (black bars) were used in experiment A, and nine F_0 , six $F_{0.25}$ and six $F_{0.375}$ lines in experiment B. There was no detected line effect on mating success (Supplementary Information). F_0 male mating success is set at 1. The total number of matings (n) is given for each experiment. G -test significance: asterisk, $P < 0.05$.

† Present address: The Galton Laboratory, University College London, 4 Stephenson Way, London NW1 2HE, UK.

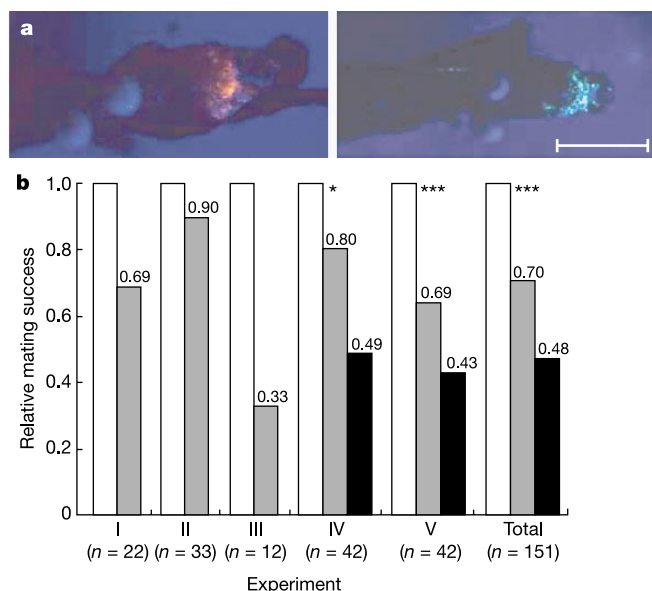


Figure 2 Mating success of inbred and outbred males in free-flight conditions. **a**, Ventral abdominal tips of mated females showing transferred orange (left) and green (right) fluorescence under ultraviolet illumination. Scale bar, ~3 mm. **b**, Mating success of inbred males relative to that of outbred males (F_0 , set at 1), standardized by the relative disappearance of males. Five replicate experiments, I to V, were performed using butterflies from 15 F_0 and 14 $F_{0.25}$ independent lines in experiments I–III, and from 15 F_0 , 12 $F_{0.25}$ and 6 $F_{0.375}$ lines in experiments IV and V. No $F_{0.375}$ males were released in experiments I–III. G -test significance: asterisk, $P < 0.05$; three asterisks, $P < 10^{-4}$. White bars, F_0 lines; grey bars, $F_{0.25}$ lines; black bars, $F_{0.375}$ lines.

involves a perch-and-chase strategy^{19,20}. Males take off from chosen perches to investigate any flying object they detect. Females are then pursued, such chases being interrupted by repeated alightings. Close-range courtship signals then come into play, involving flickering of the male's wings and the release of pheromones from exposed androconia, followed by attempts of the male to latch onto the female's abdomen¹⁹. In small cages, the opportunity for 'normal' mate detection and chase are effectively absent; courtship can often be initiated by chance encounters, and females might also be unable to reject copulation attempts.

We therefore performed large-scale experiments in semi-natural conditions, allowing free flight and a fuller expression of mate location and courtship behaviour, by releasing temporary populations of butterflies in a spacious tropical greenhouse. The butterflies behaved naturally in a vegetation structure and at an adult density comparable to many local field populations²⁰; perching behaviour and male–male interactions seemed to us closely comparable to those in the field. Equal numbers of males of differing inbreeding coefficients were used in each experiment. Males had their genitalia dusted with fluorescent dusts, a different colour being used for each inbreeding level, before being released and allowed to establish themselves in the vegetation of the greenhouse. They then competed over a period of 36–48 h for about half as many virgin females. Dust was transferred to the female's abdomen during mating; after recapture this enabled the identification of the class of male with which each female had mated (Fig. 2a, Table 1) (occasional double matings were detected). Again, outbred males showed a higher rate of mating than either class of inbred male (F_0 : $F_{0.25}$ comparison, $G = 8.91$, d.f. = 1, $P < 2 \times 10^{-3}$; F_0 : $F_{0.375}$ comparison, $G = 18.6$, d.f. = 1, $P < 2 \times 10^{-5}$). Highly inbred males achieved only 14% of all matings, representing a mating success about one-half that of outbred males. Males of the intermediate level of inbreeding had an intermediate success equivalent to 70% of outbred males (Fig. 2b; overall F_0 : $F_{0.25}$: $F_{0.375}$ comparison, $G = 21.36$, d.f. = 2, $P < 2 \times 10^{-5}$). More interestingly, the deleterious effect of inbreeding on mating success was, overall, significantly stronger in free-flight conditions than in the laboratory ($\chi^2 = 9.36$, d.f. = 2, $P < 0.01$), strongly suggesting that the cages allowed only a partial estimation of inbreeding depression.

Outbred males were recaptured in higher numbers than inbred males at the end of the experiments (Table 1). Although the difference is not significant, the trend could contribute to the decreased net mating success of inbred males. Nevertheless, inbred butterflies still mated significantly less often when the relative loss of males was controlled for (pooled data: $G = 8.52$, d.f. = 2, $P = 0.014$). Thus, the slightly faster disappearance of inbred males during the experiments (due to effects on longevity^{9,10,15,16}, escape or lethargy^{16,18}) cannot fully explain the decreased mating success in free-flying *B. anymana*; at least 60% of the estimated genetic load can be accounted for by differences in mating ability, excluding survivorship (Table 2). Inbreeding must therefore affect other traits that are expressed, or detected, only in free-flying conditions. The genetic load estimated for male mating success in the greenhouse (Table 2) is comparable to that for egg hatching (2.70 lethal equivalents per gamete in ref. 10).

Table 1 Female mating and male recapture in greenhouse experiments

Parameter	Experiment					
	I	II	III	IV	V	Total
Test	F_0 : $F_{0.25}$	F_0 : $F_{0.25}$	F_0 : $F_{0.25}$	F_0 : $F_{0.25}$: $F_{0.375}$	F_0 : $F_{0.25}$: $F_{0.375}$	F_0 : $F_{0.25}$: $F_{0.375}$
Dust colour	O:G	O:G	O:G	G:O:Y	G:O:Y	
Females (F_0)						
Released	50	72	50	90	90	352
Recaptured (%)	44.0	48.6	24.0	56.7	56.7	48.6
Mated with						
F_0	13	17	9	20	25	84
$F_{0.25}$	6	14	3	14	12	49
$F_{0.375}$	–	–	–	7	5	12
$F_0 + F_{0.25}$	3	2	0	1	0	6
Unmated	0	2	0	9	9	20
Males						
Released	47:47	59:59	48:48	60:60:60	60:60:60	274:274:120
Recaptured (%)	44.7	44.1	29.2	47.2	34.4	40.3
F_0	24	27	14	33	28	126
$F_{0.25}$	18	25	14	29	21	107
$F_{0.375}$	–	–	–	23	13	36

For males, the figures are absolute numbers after recapture for each inbreeding coefficient. For females (all $F = 0$), the figures are the numbers of recaptured females that mated with each F -class of male. Dust colours were orange (O), green (G) and yellow (Y) (see Fig. 2a). Overall recapture rates (%) are given for each experiment.

Table 2 Genetic load for mating success in caged and free-flying male butterflies

Experiment	Comparison	Genetic load	s.e.	P
Trials in captivity	$F_0:F_{0.25}$	1.50 (1.17, 1.83)	0.07	0.002
	$F_0:F_{0.25}:F_{0.375}$	0.83 (-0.03, 1.70)	0.31	0.055
Free-flight experiments	$F_0:F_{0.25}$ (I–II)	2.58 (0.41, 4.76)	0.78	0.029
		2.10 (-0.71, 4.91)	1.01	0.107
	$F_0:F_{0.25}:F_{0.375}$ (IV, V)	3.37 (1.53, 5.21)	0.66	0.006
		1.96 (0.57, 3.36)	0.50	0.017
	Total (I–V)	2.75 (0.92, 4.59)	0.82	0.007
		1.73 (0.49, 2.98)	0.55	0.010

Lethal equivalents per gamete are estimated from the slope of the regression of the natural-log-transformed mating success on the inbreeding coefficient, for the given classes of inbreeding and experiments^{10,30}. The mating success in cage experiments is calculated on the basis of the proportion of matings in each replicate experiment. The mating success in greenhouse (free-flight) experiments is calculated as the number of matings divided (in the upper line of each pair) by the number of males at release or (in the lower line) after recapture. Lower and upper 95% confidence interval boundaries are given in parentheses. Note the wider confidence intervals for greenhouse estimates; these are likely to result from additional sources of variation, such as the amount of sunshine, inherent in more natural experimental conditions.

The marked decrease in mating success in male *B. anynana* due to inbreeding could have different causes. Disruption of close-range courtship signals (such as pheromone profiles) could act both in captivity and in the semi-natural conditions. However, other components of male mating success are likely to be expressed only or mainly in free-flying populations. Locomotor or athletic abilities strongly affect male mating success in some inbred *Drosophila* lines^{16,18}, and these effects could also be important in our greenhouse experiments. They could, for example, affect the alertness and responsiveness of perching males, their persistence in chasing females once courtship is initiated, and the outcome of male–male interactions, which are all expected to lead to an overall lower mating of inbred males. The apparent genotype–environment interaction results mainly from the poor performance of the most inbred males in the greenhouse (Figs 1 and 2). Although the use of more classes of inbreeding would be needed to confirm this interaction, it could result from different behavioural characters showing inbreeding depression at different levels of inbreeding in combination with conditions under which characters are expressed differentially^{3,7}.

Inbreeding can influence the survival of populations^{2,3}, for example through a direct effect on individual survival⁴. The poor mating vigour of inbred males in *B. anynana* indicates that the fragmentation of populations with a high genetic load⁶ might lead to a general decrease in mating activity. Outbred immigrants into small, isolated populations would then enjoy an immediate mating advantage, thereby enhancing the heterosis-assisted^{21,22} restoration of mean population fitness unless offset by ecological effects or outbreeding depression²³. If, as is likely, long-distance migrants are more outbred, this process could be important in the frequency-dependent maintenance of genetic diversity at a regional scale in fragmented landscapes²⁴.

Our results highlight the necessity for a field validation of behavioural data from captive animals, especially those involving interactions and signalling, as in mate choice experiments (such as ref. 19). Captivity constrains behavioural responses and may truncate the expression of key courtship characters. Similarly, estimates of inbreeding depression in invertebrates are taken almost exclusively from laboratory experiments^{1,3,6,16}. However, these are likely to be severe underestimates, particularly when stressful field conditions^{15,25} or relaxed constraints on the expression of behavioural traits⁹, as in our study, magnify inbreeding depression (but see ref. 8 for a case in which it is not magnified). A low mating success of inbred males, as occurred in our greenhouse experiments, could directly affect the success of reintroduction programmes from captive stock²⁶. Conserving genetic diversity through such programmes might therefore be a risky enterprise unless backed up by observations from the wild¹².

The technique we have developed here for comparing the mating success of different cohorts of animals is widely applicable to many issues in sexual selection and mate choice^{19,27}, and could be used in natural environments. It could also be combined with additional studies for key individuals, for example to examine variation in mating success within cohorts²⁸, or to determine the outcome of sperm competition for double-mated females²⁷. Such approaches would reveal more about the proximal reasons for differences in mating patterns and reproductive success, as well as the consequences of these differences. □

Methods

Inbred and outbred lines

An outbred stock of *B. anynana* from Malawi has been maintained with high genetic diversity in the laboratory for more than 100 generations; the adult population size is several hundred individuals with an effective population size ~60% of the total population²⁸. Mating pairs were drawn from this stock and their progeny were reared at 27 °C, giving independent lines with an inbreeding coefficient $F \approx 0$. From these F_0 lines, brother–sister mating pairs were drawn, giving lines with $F = 0.25$. Again, from these $F_{0.25}$ lines, brother–sister mating pairs were drawn, giving lines with $F = 0.375$. Independent F_0 , $F_{0.25}$ and $F_{0.375}$ lines were available simultaneously after three generations; all replicate experiments involved, for each class of males, individuals taken from 6–15 synchronous independent lines (see Figs 1 and 2; Supplementary Information). Larvae were reared with an excess of food in sleeves containing about 30 larvae each; inbred and outbred lines were interspersed in the rearing chamber. Rearing details are described elsewhere²⁸.

Cage trials

Each trial involved competition between three males, one of each inbreeding class, for a single unrelated, outbred female, in a $12 \times 20 \times 28 \text{ cm}^3$ mesh cage, at 27 °C. Virgin males 3–5 days old were introduced into each cage in the afternoon and left to settle overnight; about 60 min after the lights had been turned on the next morning, a single 2-day-old virgin outbred female ($F = 0$) was introduced, followed by monitoring of matings every 5–10 min. Mating in *B. anynana* lasts on average ~30 min (Supplementary Information). Butterflies from cages in which mating had occurred were not used again. If no mating had occurred after 120 min, the unmated female was discarded, whereas males were fed and tested on the following day with a new virgin female (but there was then no further re-use of males). Two replicate experiments were conducted with different sets of lines: 76 and 109 matings were recorded in 100 and 120 trials, respectively (Fig. 1).

Free-flight experiments

We treated the genitalia of 3–7-day-old virgin males with coloured 'rodent-tracking' fluorescent dust, using a different colour for each inbreeding class. Males were dusted on the same morning and released at midday in a greenhouse that provided a $15 \times 15\text{-m}^2$ flight area with a temperature of 24–32 °C, high humidity and tropical vegetation. A high and more open space covered a central, circular pond around which rotten fruit was placed as a food source to give an environment comparable to local habitat patches characterized by a high adult density in the wild²⁰. Males were left to adapt and interact with one another. On the following morning, 2–5-day-old outbred ($F = 0$) virgin females were released, in numbers corresponding to a ~3:1 male:female ratio at release (Table 1; using $F = 0.25$ females gave similar results; see Supplementary Information). A second group of females, about half the size of the first group, was released 1 day later. All butterflies were netted 1.5 days after the last group of females was released, until no more butterflies were found. This procedure resulted in a high mating rate of recaptured females (88%) while ensuring a relatively high competition for pairings between males. Note that virgin females are very rare in the field²⁰, and the ratio of males to receptive females used here is probably substantially lower than occurs in nature. Dust is consistently transferred to the female's genitalia and abdominal folds during mating, and pilot experiments showed no effect of dust-colour on mating success (Supplementary Information). As an internal control for any possible effect, we switched colours used for F_0 and $F_{0.25}$ males between experiments (Table 1). Recaptured females were inspected for fluorescence with a binocular microscope under ultraviolet illumination to record the class of male(s) they had mated with. Colours were chosen so that double fluorescence could be detected and scored unequivocally. Recaptured males were also scored to monitor survivorship. Five experiments were performed sequentially (Table 1).

Statistical analyses

Significance was tested by using G-tests²⁹. Laboratory mating distributions were compared with 1:1:1 or 1:(1 + 1) distributions. For free-flight experiments, replicates I–III were compared with a 1:1 distribution (two classes of males), and replicates IV and V with a 1:1:1 distribution (three classes). Detected double matings (that is, those involving males of differing inbreeding coefficients) were scored as 0.5 for each class of male. Although all replicates involved males in even numbers, the total is uneven because $F_{0.375}$ males were not always used. The overall distribution of matings was therefore compared with the actual distribution of released males: 274:274:120 ($F_0:F_{0.25}:F_{0.375}$). To correct for male disappearance, we compared the distribution of matings with the distribution of males after recapture, for example 126:107:36 ($F_0:F_{0.25}:F_{0.375}$) for the overall data set (Table 1).

We tested for a stronger effect of inbreeding in the greenhouse versus cages by using a 2×3 (treatments $\times$ inbreeding classes) heterogeneity test with a Williams' correction²⁹ (experiments IV and V only).

Received 27 January; accepted 24 April 2003; doi:10.1038/nature01713.

- Keller, L. F. & Waller, D. M. Inbreeding effects in the wild. *Trends Ecol. Evol.* **17**, 230–241 (2002).
- Saccheri, I. *et al.* Inbreeding and extinction in a butterfly metapopulation. *Nature* **392**, 491–494 (1998).
- Bijsma, R., Bundgaard, J. & Boerema, A. C. Does inbreeding affect the extinction risk of small populations? Predictions from *Drosophila*. *J. Evol. Biol.* **13**, 502–514 (2000).
- Keller, L. F., Arcese, P., Smith, J. N. M., Hochachka, W. M. & Stearns, S. C. Selection against inbred song sparrows during a natural population bottleneck. *Nature* **372**, 356–357 (1994).
- Roff, D. A. & DeRose, M. A. The evolution of trade-offs: Effects of inbreeding on fecundity relationships in the cricket *Gryllus firmus*. *Evolution* **55**, 111–121 (2001).
- Saccheri, I., Brakefield, P. M. & Nichols, R. A. Severe inbreeding depression and rapid fitness rebound in the butterfly *Bicyclus anynana* (Satyridae). *Evolution* **50**, 2000–2013 (1996).
- DeRose, M. A. & Roff, D. A. A comparison of inbreeding depression in life-history and morphological traits in animals. *Evolution* **53**, 1288–1292 (1999).
- Armbruster, P., Hutchinson, R. A. & Linzell, T. Equivalent inbreeding depression under laboratory and field conditions in a tree-hole-breeding mosquito. *Proc. R. Soc. Lond. B* **267**, 1939–1945 (2000).
- Meagher, S., Penn, D. J. & Potts, W. K. Male–male competition magnifies inbreeding depression in wild house mice. *Proc. Natl Acad. Sci. USA* **97**, 3324–3329 (2000).
- van Oosterhout, C., Zijlstra, W. G., van Heuven, M. K. & Brakefield, P. M. Inbreeding depression and genetic load in laboratory metapopulations of the butterfly *Bicyclus anynana*. *Evolution* **54**, 218–225 (2000).
- Nieminen, M., Singer, M. C., Fortelius, W., Schöps, K. & Hanski, I. Experimental confirmation that inbreeding depression increases extinction risk in butterfly populations. *Am. Nat.* **157**, 237–244 (2001).
- Madsen, T., Shine, R., Olsson, M. & Wittzell, H. Restoration of an inbred adder population. *Nature* **402**, 34–35 (1999).
- Coulson, T. N. *et al.* Genotype by environment interactions in winter survival in red deer. *J. Anim. Ecol.* **67**, 434–445 (1998).
- Höglund, J. *et al.* Inbreeding depression and male fitness in black grouse. *Proc. R. Soc. Lond. B* **269**, 711–715 (2002).
- Jimenez, J. A., Hughes, K. A., Alaks, G., Graham, L. & Lacy, R. C. An experimental study of inbreeding depression in a natural habitat. *Science* **266**, 271–273 (1994).
- Miller, P. S., Glasner, J. & Hedrick, P. W. Inbreeding depression and male-mating behavior in *Drosophila melanogaster*. *Genetica* **88**, 29–36 (1993).
- Meffert, L. M. & Bryant, E. H. Mating propensity and courtship behavior in serially bottlenecked lines of the housefly. *Evolution* **45**, 293–306 (1991).
- Maynard Smith, J. Fertility, mating behaviour, and sexual selection in *Drosophila subobscura*. *J. Genet.* **54**, 261–279 (1956).
- Breuker, C. J. & Brakefield, P. M. Female choice depends on size but not symmetry of dorsal eyespots in the butterfly *Bicyclus anynana*. *Proc. R. Soc. Lond. B* **269**, 1233–1239 (2002).
- Brakefield, P. M. & Reitsma, N. Phenotypic plasticity, seasonal climate and the population biology of *Bicyclus* butterflies (Satyridae) in Malawi. *Ecol. Entomol.* **16**, 291–303 (1991).
- Saccheri, I. & Brakefield, P. M. Rapid spread of immigrant genomes into inbred populations. *Proc. R. Soc. Lond. B* **269**, 1073–1078 (2002).
- Ebert, D. *et al.* A selective advantage to immigrant genes in a *Daphnia* metapopulation. *Science* **295**, 485–488 (2002).
- Marr, A. B., Keller, L. F. & Arcese, P. Heterosis and outbreeding depression in descendants of natural immigrants to an inbred population of song sparrows (*Melospiza melodia*). *Evolution* **56**, 131–142 (2002).
- Hanski, I. & Gilpin, M. *Metapopulation Biology: Ecology, Genetics and Evolution* (Academic, London, 1996).
- Hedrick, P. W. & Kalinowski, S. T. Inbreeding depression in conservation biology. *Annu. Rev. Ecol. Syst.* **31**, 139–162 (2000).
- Miller, P. S. & Hedrick, P. W. Inbreeding and fitness in captive populations—lessons from *Drosophila*. *Zoo Biol.* **12**, 333–351 (1993).
- Tregenza, T. & Wedell, N. Polyandrous females avoid the costs of inbreeding. *Nature* **415**, 71–73 (2002).
- Brakefield, P. M. *et al.* Effective population size, reproductive success and sperm precedence in the butterfly *Bicyclus anynana* in captivity. *J. Evol. Biol.* **14**, 148–156 (2001).
- Sokal, R. R. & Rohlf, F. J. *Biometry: The Principles and Practice of Statistics in Biological Research* (W. H. Freeman, New York, 1995).
- Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits* (Sinauer, Sunderland, Massachusetts, 1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank I. Saccheri, B. Zwaan and C. Breuker for help with the experimental design and preliminary experiments, M. Begon for providing fluorescent dusts, and N. Wurzer and M. Lavrijsen for technical help and host-plant cultivation. We are grateful to the Hortus Botanicus Leiden for access to the greenhouse. Finance and support was provided by the European Union TMR Network 'Survival of Species in Fragmented Landscapes' (FRAGLAND).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.J. (m.joron@ucl.ac.uk) or P.M.B. (brakefield@rulsfb.leidenuniv.nl).

Dosage sensitivity and the evolution of gene families in yeast

Balázs Papp*†, Csaba Pál*† & Laurence D. Hurst*

* Department of Biology and Biochemistry, University of Bath, Bath, Somerset BA2 7AY, UK

† Department of Plant Taxonomy and Ecology, Eötvös Loránd University, Pázmány Péter Sétány 1/C, H-1117 Budapest, Hungary

According to what we term the balance hypothesis, an imbalance in the concentration of the subcomponents of a protein–protein complex can be deleterious¹. If so, there are two consequences: first, both underexpression and overexpression of protein complex subunits should lower fitness, and second, the accuracy of transcriptional co-regulation of subunits should reflect the deleterious consequences of imbalance. Here we show that all these predictions are upheld in yeast (*Saccharomyces cerevisiae*). This supports the hypothesis^{2,3} that dominance is a by-product of physiology and metabolism rather than the result of selection to mask the deleterious effects of mutations. Beyond this, single-gene duplication of protein subunits is expected to be harmful, as this, too, leads to imbalance. As then expected, we find that members of large gene families are rarely involved in complexes. The balance hypothesis therefore provides a single theoretical framework for understanding components both of dominance and of gene family size.

About 30% of yeast genes code for proteins that are involved in annotated (experimentally confirmed) protein complexes⁴. Consider a complex formed by the binding of proteins A and B. There are numerous reasons¹ why an excess of A, for example, might be deleterious: A could form homodimers with a function different from that of the AB heterodimer⁵, it might be a regulatory subunit that competes with other regulatory subunits to bind the catalytic subunit B (ref. 6), it might be toxic by binding irreversibly to targets where AB should bind normally⁷, or it could form toxic precipitates⁸. Additionally, subunits forming a bridge between parts of a complex can inhibit complex assembly if present in excess^{1,9} (Supplementary Information).

If imbalance were deleterious¹ (the balance hypothesis) we would expect adaptations to minimize the degree of imbalance. Rapid degradation of unassembled ribosomal subunits¹⁰ is likely to be one of these. The balance hypothesis also predicts that a greater decrease in fitness should be seen in cells that are heterozygotes for knockouts of single genes if the gene is involved in a complex than if it is not. A systematic mutagenesis experiment¹¹ allows comparison of the dosage sensitivity of many genes. For nearly all single-gene deletions in the yeast genome the growth rates of heterozygous and homozygous diploid strains are known¹¹. To minimize any measurement biases we consider only essential genes (lethal homozygote deletion). The decrease in mean fitness of heterozygotes compared with the wild type is 5%, and only a few knockouts in essential genes have a large effect on fitness (for distribution see Supplementary Information). To test whether dosage-sensitive genes are more likely to be involved in protein complexes, we used an annotated list of known complexes in yeast⁴. Unfortunately, the list is not complete and might be biased, so an extended set of protein interactions was also used (Supplementary Information).

As predicted, genes with low heterozygote fitness tend to be in complexes: genes with less than 5% fitness deficiency constitute 52% of the 816 proteins investigated, but only 37% of them are involved in protein complexes, whereas of those with high fitness deficiency (more than 15%), more than 88% of them are known to interact with other proteins (Fig. 1). This implies that dosage-sensitive genes are at least twice as likely to be involved in protein

complexes than genes with low dosage sensitivity ($\chi^2 = 19.78$, d.f. = 1, $P < 10^{-5}$). The result remains even when only those genes with catalytic activity are considered: enzymes that act in a complex have smaller heterozygote fitness than the rest (Mann–Whitney U -test, $P < 0.01$). We also observe that complex size has some influence on dosage sensitivity: when only genes involved in one protein complex are considered, heterozygote fitness shows a weak negative association with the size of the complex (Spearman $\rho = -0.24$, $P < 10^{-3}$, $N = 207$). The number of protein complexes that a gene is involved in has no detectable effect on heterozygote fitness (Spearman $\rho = -0.067$, $P = 0.27$). Being in a complex therefore seems to be the more important factor affecting dosage sensitivity (see Supplementary Information).

These results, although consistent with the balance hypothesis, might also reflect the fact that the concentration of the complex decreases sharply with decreased gene dosage, as interacting proteins have a small chance to assemble^{1,12}. In this alternative model we do not expect that an artificially increased dosage of one subunit need have any deleterious effect, as long as the dosage of the complex is unaffected. Can we discriminate between this, the complex concentration model, and the balance hypothesis? The balance hypothesis has three unique predictions: first, artificial overexpression of one subunit should be deleterious; second, the strength of transcriptional co-regulation of subunits is expected to reflect dosage sensitivity; last, single-gene duplication of subunits should be harmful, as this can lead to an immediate imbalance of constituents. All three predictions are upheld in yeast.

According to the balance hypothesis, deficiencies caused by halving gene dosage in one gene are rescued, at least in part, by decreasing expression in the interacting partner. In yeast, there is experimental evidence that suppression of mutation in one gene is sometimes achieved by a decreased gene dosage of another gene¹³. The balance hypothesis also correctly predicts that if overexpression of one subunit is deleterious, then overproduction of its stoichiometric partner should rescue the cell^{7,13}.

Do we find more generally that artificially induced overexpression tends to be especially harmful for proteins involved in complexes? Using a publicly available database⁴, we compiled a data set of proteins for which artificial overexpression of the gene in wild-type cells is either lethal ($N = 36$) or has no clear detrimental effect on fitness ($N = 100$, see Methods). Of the genes in the former group, 47% are members of a protein complex. This is a highly significant excess compared with the rest of the data set, in which

only 8% of genes participate in complexes ($\chi^2 = 27.14$, d.f. = 1, $P < 10^{-6}$).

The balance hypothesis also predicts that precise transcriptional co-regulation of interacting proteins should be especially important when the cell is sensitive to changes in the ratio of the subunits. Indeed, interacting proteins are more frequently co-expressed than random pairs¹⁴. Might it also be that dosage sensitivity of subunits affects the strength of co-regulation? To test this idea we calculated the mean of heterozygote fitnesses for each pair of interacting proteins (ignoring duplicates). We also derived a correlation of expression patterns across time for each gene pair (resulting in 454 pairs; see Methods). In agreement with expectations, the frequency of co-expressed gene pairs sharply falls with increasing mean heterozygote fitness (Fig. 2). In particular, only 20% of the interacting pairs with less than 5% fitness deficiency show evidence for co-expression. By contrast, of the subunit pairs with relatively high fitness deficiency (more than 15%), more than 80% of them are co-expressed. (The above result does not contradict a report of there being no relationship between change in mRNA level and the fitness of homozygote knockouts¹⁵, not least because we investigated the association between the strength of co-expression, not the rate of expression, and fitness.)

The above results support the notion that dominance might be a by-product of physiological mechanisms^{2,3} rather than an adaptation to shield the harmful effects of mutations¹⁶. Beyond this, the balance hypothesis has ramifications for understanding variation in gene family size. If an imbalance is deleterious, then single-gene duplications, increasing protein dosage, should be more likely to be counter-selected if the gene product is involved in complexes. By contrast, duplication by genome duplication increases the dosage of all genes and should not affect the balance. We therefore expect the two sorts of duplication event to have different consequences. We observe that genes involved in protein complexes rarely have numerous paralogues in the yeast genome. Whereas 33% of the genes without any paralogue participate in protein complexes, this frequency drops gradually to about 21% for genes with three or more paralogues ($N = 3326$, $\chi^2 = 36.52$, d.f. = 3, $P < 10^{-7}$; see Methods).

Furthermore, the balance hypothesis uniquely predicts the co-evolution of protein subunits: either interacting gene pairs should remain solo copies or both should undergo gene duplication. Indeed, if we consider interacting subunit pairs, we find a large excess of interacting pairs with the same number of paralogues (we

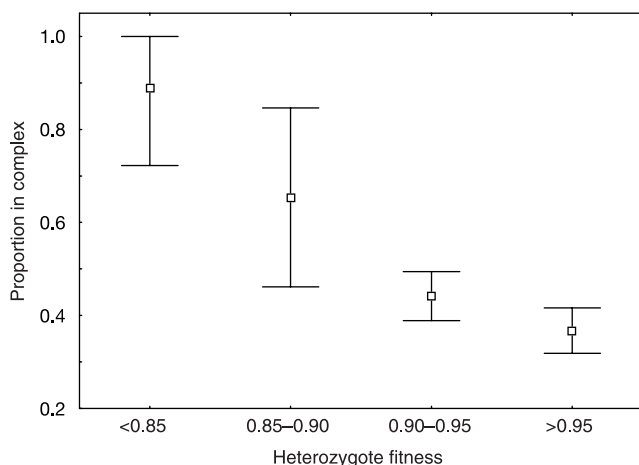


Figure 1 Proportion of genes in protein complex as a function of heterozygote fitness. Only essential genes are considered. Confidence intervals (95%) were obtained by the bootstrap method³⁰.

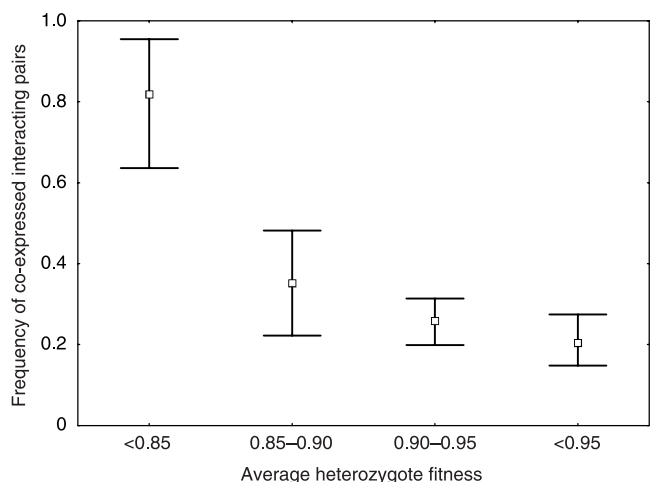


Figure 2 The frequency of co-expressed interacting gene pairs decreases with increasing mean heterozygote fitness ($N = 454$, $\chi^2 = 37.59$, d.f. = 3, $P < 10^{-7}$). Confidence intervals (95%) were obtained by the bootstrap method³⁰.

observe 4,321 of 6,927 pairs, but only about 2,965 are expected by chance; $P < 10^{-5}$ from randomization; see Methods). There is also a large excess of solo copy pairs (1,541 observed but only about 878 are expected by chance; $P < 10^{-5}$).

After polyploidization the loss of duplicate copies of interacting genes one at a time leads to imbalance. The balance hypothesis therefore predicts which functional classes have a tendency to remain duplicated after a genome duplication. For example, ribosomes are needed at a high titre, but an imbalance in the concentration of ribosomal proteins is harmful¹⁷. We therefore expect ribosomal genes to be over-represented in the class of duplicates derived from whole-genome duplication. To investigate this, we compiled a set of genes with only one paralogue in the yeast genome (see Methods), resulting in 1,020 genes from which 393 genes are putatively derived from whole-genome duplication¹⁸ (although we cannot be certain about the classification). The frequency of cytosolic ribosomal genes (20.61%) is much higher among whole-genome duplicates than in single-gene duplicates (2.39%) ($\chi^2 = 94.04$, d.f. = 1, $P < 10^{-21}$).

In mammals¹⁹ and yeast⁷, specialized chaperones protect the cell from free excess subunits of certain protein complexes. Although this indicates that the problems of imbalance might not be restricted to yeast, we do not wish to suggest that the balance of proteins in complexes is the only concern about dominance and gene family size in this or other species. For example, the lack of elaborate developmental pathways in yeast compared with vertebrates will probably ensure that changes in dosage of transcription factors will have different phenotypic consequences in the two lineages. Most genes in yeast influence the expression of relatively few other genes²⁰; this contrasts with the long regulatory cascades during multicellular development²¹, in which slight perturbations can have large effects. Pivotal developmental regulators in *Drosophila* and *Caenorhabditis* are also particularly likely to be haploinsufficient²². By contrast, this functional group is not an outlier in yeast (data not shown). We therefore expect there to be species differences in what might be the most important genes as regards perturbation of dosage. Moreover, although—as we have shown—not all genes are equally free to duplicate, adaptation must explain some of the variation in gene family size (for example the expansion of olfactory system genes in mammals²³). Nevertheless, we expect that protein–protein interactions influence both dominance and the size of gene families in all species. □

Methods

Fitness calculation

The growth rates of heterozygous and homozygous strains were as measured in ref. 11. We used the growth rates obtained on YPD (fermentable) substrate because yeast preferentially ferments glucose²⁴. Only genes with two measurements from repeat experiments were retained, and average growth rates (R_{av}) were calculated. Relative heterozygote fitness was calculated as $F_{het} = R_{av}/R_{max}$ where R_{max} is the maximal average growth rate of the knockouts. We also repeated all analyses using raw, rather than normalized, growth rate but observed no qualitative changes in the results. Overlapping genes, being indistinguishable by deletion analysis, were excluded.

Let the fitness of the wild type, heterozygote and homozygote knockouts be 1, $1 - h$ and $1 - s$, respectively, where h denotes the dominance of heterozygote mutations. Only essential genes ($s = 1$) were considered, so heterozygote fitnesses ($1 - h$) are proportional to dominance. Classification of the dispensability of genes (essential versus non-essential) was provided by the Yeast Proteome Database²⁵. Only essential genes that also failed to grow in the homozygous knockout condition in ref. 11 were analysed further.

Physical interactions and overexpression phenotypes

A set of annotated protein complexes was assembled from the MIPS Comprehensive Yeast Genome Database (CYGD) catalogue of known protein complexes⁴. Where annotation was hierarchical, the lowest level of hierarchy was used. The composition of each complex was verified manually to discriminate between alternative subtypes of the same complex. Pairwise interactions were derived by considering all possible protein pairings within complexes (yielding 7,571 pairs).

To minimize any potential annotation bias, we extended our analysis to include protein complexes detected by the high-throughput TAP procedure²⁶. All our analyses were repeated on this larger data set and yielded similar results (Supplementary Information). Owing to high false-positive error rates²⁷, we did not include yeast two-hybrid data. The

TAP procedure²⁶ is an effective approach for identifying stable protein interactions, but nevertheless suffers from a ~30% false-positive rate and currently covers only ~25% of the yeast proteome²⁶.

Data on artificial overexpression studies were obtained from the MIPS CYGD⁴ database. Genes for which a lethal or toxic overexpression phenotype was reported in an otherwise wild-type background were classified as lethal ($N = 36$). This group was compared with a group of genes with no clear detrimental phenotypic effect when overexpressed in wild types ($N = 100$).

Sequence and expression similarity of protein subunits

Sequence similarity of interacting pairs was computed with a pairwise BLASTP search²⁸ (pairs above a conservative cut-off of $E = 10^{-2}$ were considered to be duplicates and removed). Expression similarities of interacting proteins were calculated with compiled whole-genome mRNA expression data (<http://rana.lbl.gov/EisenData.htm>). The data set contains information from 80 independent microarray experiments in which changes in expression levels were calculated at multiple time points, including cell cycle, sporulation and diauxic shift (see ref. 29). For each gene and each time point the data set provides normalized ratios of expression level in experimental and reference populations of cells. For the 454 pairs of genes with at least 70 time points, the Pearson correlation coefficient (r) of expression ratios was calculated. If r was higher than a cut-off value (r_0), then the pair was considered to be co-expressed. r_0 was calculated from the correlation coefficient of 10,000 randomly selected gene pairs, using the 95% confidence interval derived from the data ($r_0 \approx 0.566$). The use of more stringent criteria for co-expression ($r_0 = 0.7$) gives the same qualitative result (see Supplementary Information).

Calculation of gene family size

A BLASTP search²⁸ of all yeast proteins against each other was performed to identify putative paralogues. The number of paralogues of a given gene was estimated by the number of its BLASTP hits with expected values 10^{-10} or less. Using different cut-offs ($E < 10^{-5}$ or $E < 10^{-2}$) gives comparable results (see Supplementary Information). When considering the relationship between gene family size and the frequency of being in complexes, we excluded duplicates left over from the genome duplication event¹⁸. As the group of unclassified proteins⁴ is strongly biased towards genes without any paralogue in the genome ($\chi^2 = 631.8$, d.f. = 1, $P < 10^{-138}$) and unclassified proteins cannot participate in annotated protein complexes by definition, we also excluded this group. None of our previous results is affected by this bias.

Randomization of protein interactions

To investigate whether interacting protein subunits are more likely to have the same number of paralogues than expected by chance, we compiled a list of interacting protein pairs with the following properties: first, they must be members of the same protein complex; second, they must show no sequence similarity to each other. From the 6,927 collected interacting gene pairs, 4,321 cases were found in which members of the pair had the same gene family size. Using this data set, we assigned the pairs randomly and recalculated the number of pairs with the same number of paralogues in the yeast genome. This procedure was repeated 10^5 times. A similar protocol was used to examine whether there is an excess of interacting solo copy pairs.

Received 7 March; accepted 13 May 2003; doi:10.1038/nature01771.

1. Veitia, R. A. Exploring the etiology of haploinsufficiency. *BioEssays* **24**, 175–184 (2002).
2. Wright, S. Physiological and evolutionary theories of dominance. *Am. Nat.* **68**, 25–53 (1934).
3. Kacser, H. & Burns, J. A. The molecular basis of dominance. *Genetics* **97**, 639–666 (1981).
4. Mewes, H. W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **30**, 31–34 (2002).
5. Rottensteiner, H., Kal, A. J., Hamilton, B., Ruis, H. & Tabak, H. F. A heterodimer of the Zn2Cys6 transcription factors Pip2p and Oaf1p controls induction of genes encoding peroxisomal proteins in *Saccharomyces cerevisiae*. *Eur. J. Biochem.* **247**, 776–783 (1997).
6. Wu, X., Hart, H., Cheng, C., Roach, P. J. & Tatchell, K. Characterization of Gac1p, a regulatory subunit of protein phosphatase type I involved in glycogen accumulation in *Saccharomyces cerevisiae*. *Mol. Genet. Genomics* **265**, 622–635 (2001).
7. Abruzzi, K. C., Smith, A., Chen, W. & Solomon, F. Protection from free beta-tubulin by the beta-tubulin binding protein Rbl2p. *Mol. Cell Biol.* **22**, 138–147 (2002).
8. Yuan, J. *et al.* The instability of the membrane skeleton in thalassemic red blood cells. *Blood* **86**, 3945–3950 (1995).
9. Bray, D. & Lay, S. Computer-based analysis of the binding steps in protein complex formation. *Proc. Natl Acad. Sci. USA* **94**, 13493–13498 (1997).
10. Warner, J. R. The economics of ribosome biosynthesis in yeast. *Trends Biochem. Sci.* **24**, 437–440 (1999).
11. Steinmetz, L. M. *et al.* Systematic screen for human disease genes in yeast. *Nature Genet.* **31**, 400–404 (2002).
12. van Holde, K. E., Johnson, W. C. & Shing Ho, P. *Principles of Physical Biochemistry* (Prentice Hall, Upper Saddle River, New Jersey, 1998).
13. Stevens, R. C. & Davis, T. N. Mlc1p is a light chain for the unconventional myosin Myo2p in *Saccharomyces cerevisiae*. *J. Cell Biol.* **142**, 711–722 (1998).
14. Ge, H., Liu, Z., Church, G. M. & Vidal, M. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nature Genet.* **29**, 482–486 (2001).
15. Gievers, G. *et al.* Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391 (2002).
16. Fisher, R. A. *The Genetical Theory of Natural Selection* (Clarendon, Oxford, 1930).
17. Li, B., Vilardell, J. & Warner, J. R. An RNA structure involved in feedback regulation of splicing and of translation is critical for biological fitness. *Proc. Natl Acad. Sci. USA* **93**, 1596–1600 (1996).
18. Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708–713 (1997).

19. Kihm, A. J. *et al.* An abundant erythroid protein that stabilizes free α -haemoglobin. *Nature* **417**, 758–763 (2002).
20. Hughes, T. R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).
21. Carroll, S. B., Grenier, J. K. & Weatherbee, S. D. *From DNA to Diversity* (Blackwell Science, Malden, Massachusetts, 2001).
22. Hodgkin, J. Fluxes, doses and poisons—molecular perspectives on dominance. *Trends Genet.* **9**, 1–2 (1993).
23. Kratz, E., Dugas, J. C. & Ngai, J. Odorant receptor gene regulation: Implications from genomic organization. *Trends Genet.* **18**, 29–34 (2002).
24. Johnston, M. Feasting, fasting and fermenting. Glucose sensing in yeast and other cells. *Trends Genet.* **15**, 29–33 (1999).
25. Hodges, P. E., McKee, A. H., Davis, B. P., Payne, W. E. & Garrels, J. I. The Yeast Proteome Database (YPD): A model for the organization and presentation of genome-wide functional data. *Nucleic Acids Res.* **27**, 69–73 (1999).
26. Gavin, A. C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
27. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417**, 399–403 (2002).
28. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
29. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
30. Sokal, R. & Rohlf, M. *Biometry* (Freeman, New York, 1995).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank C. Scharfe, L. Steinmetz, D. Bray and B. Charlesworth for comments on the manuscript. B.P. is supported by an EU Marie Curie Fellowship, C.P. by a Royal Society/Nato Fellowship and L.D.H. by the BBSRC.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to L.D.H. (bssldh@bath.ac.uk).

Widespread horizontal transfer of mitochondrial genes in flowering plants

Ulfar Bergthorsson*, Keith L. Adams*†, Brendan Thomason*‡ & Jeffrey D. Palmer*

* Department of Biology, Indiana University, Bloomington, Indiana 47405, USA

† Department of Botany, Iowa State University, Ames, Iowa 50011, USA

‡ Department of Microbiology and Immunology, University of Michigan, Ann Arbor, Michigan 48109, USA

Horizontal gene transfer—the exchange of genes across mating barriers—is recognized as a major force in bacterial evolution^{1,2}. However, in eukaryotes it is prevalent only in certain phagotrophic protists and limited largely to the ancient acquisition of bacterial genes^{3–5}. Although the human genome was initially reported⁶ to contain over 100 genes acquired during vertebrate evolution from bacteria, this claim was immediately and repeatedly rebutted^{7,8}. Moreover, horizontal transfer is unknown within the evolution of animals, plants and fungi except in the special context of mobile genetic elements^{9–12}. Here we show, however, that standard mitochondrial genes, encoding ribosomal and respiratory proteins, are subject to evolutionarily frequent horizontal transfer between distantly related flowering plants. These transfers have created a variety of genomic outcomes, including gene duplication, recapture of genes lost through transfer to the nucleus, and chimaeric, half-monocot, half-dicot genes. These results imply the existence of mechanisms for the delivery of DNA between unrelated plants, indicate that horizontal transfer is also a force in plant nuclear genomes, and are

discussed in the contexts of plant molecular phylogeny and genetically modified plants.

We first suspected that there is horizontal transfer of mitochondrial genes by finding three striking distributional anomalies in a survey of mitochondrial gene content in angiosperms¹³. Two ribosomal protein genes, *rps2* and *rps11*, were inferred¹³ from blot hybridization data to be absent from mitochondrial DNA of all members of a vast eudicot clade comprising, respectively, 180 and 182 of the 280 angiosperms examined, with the exception of one or two highly derived members of this clade (Fig. 1). Three biological models could account for these anomalies. Two models involve the loss of each gene from mitochondrial DNA early in eudicot evolution and their subsequent re-acquisition by mitochondrial DNA much later (Fig. 1), either, by horizontal gene transfer (HGT) from some unrelated plant or, by vertical transmission, by means of intracellular gene transfer (IGT) from the nucleus of the same plant lineage. A third alternative, that these genes could have been transmitted strictly vertically and exclusively through mitochondrial DNA, would mean extraordinarily frequent and pervasive mitochondrial loss throughout all other eudicot clades in which the three ‘special retention’ cases shown in Fig. 1 are phylogenetically embedded.

To distinguish between these three possibilities, we analysed levels of sequence divergence and the phylogenetic position of 31 *rps2* and 44 *rps11* genes from a broad array of angiosperms, including the three anomalous plants and their close relatives. All three sets of anomalous genes should, if they are the product of vertical transmission (by the second or third models), group in phylogenetic trees with basal eudicots that never lost these genes from their mitochondrial genomes. Instead, however, *rps2* from *Actinidia* (kiwifruit) groups with monocot *rps2* sequences with high support (Fig. 2a). This placement strongly indicates an HGT event from monocots to eudicots.

The *rps11* genes of *Lonicera* (honeysuckle; Fig. 1a) and other Caprifoliaceae (order Dipsacales) also fail to group in the position expected for vertical transmission, nesting instead within the unrelated order (Ranunculales) with strong support from bayesian analysis and alternative topology tests (see Fig. 2b, Methods and Supplementary Information). Important additional evidence for *rps11* HGT from Ranunculales to Caprifoliaceae comes from a non-coding sequence immediately upstream of *rps11*. The two Caprifoliaceae upstream sequences cluster strongly with the *Berberis* (Ranunculales) sequence in phylogenetic trees to the exclusion of Trochodendraceae (Fig. 2c), the position expected if vertically transmitted.

The phylogenetic position of *rps11* sequences from the third anomalous group, *Betula* (birch; Fig. 1b) and other Betulaceae, is unresolved and is indeed consistent with vertical transmission (Fig. 2b). The phylogenetic evidence for recapture of *rps11* in Betulaceae therefore rests on the phylogenetically anomalous presence of *rps11* in mitochondrial DNA in this family, together with the evidence that both other such anomalies are very likely to reflect gene recapture. Analysis of sequence divergence levels provides important evidence that the putatively recaptured *rps11* gene of Betulaceae is the result of HGT rather than IGT from nucleus to mitochondrion (and further supports a horizontal origin of the *Actinidia rps2* and Caprifoliaceae *rps11* genes). Nuclear substitution rates are far higher than mitochondrial rates in angiosperms^{14,15}, such that nuclear genes of mitochondrial origin quickly become long branches in mitochondrial gene trees (refs 15 and 16, and Supplementary Fig. 1). Reverse IGT (the second model) therefore predicts a highly divergent mitochondrial *rps11* or *rps2* gene in each plant group. This is clearly not so (Fig. 2a, b, and Supplementary Fig. 1), and thus mitochondrial HGT is the best explanation.

The *rps11* phylogeny serendipitously revealed a fourth, quite remarkable and well-supported case of HGT. Phylogenetic analysis

of full-length *rps11* placed *Sanguinaria canadensis* (bloodroot; Papaveraceae), a basal eudicot, in a basal position of the monocot *rps11* clade with high support (data not shown). On closer examination, *Sanguinaria rps11* turns out to be chimaeric: its 5' half is of expected eudicot, vertical origin (Fig. 2d), but its 3' half is indisputably of monocot, horizontal origin (Fig. 2e). A test of recombination¹⁷, using *Bocconia* and *Disporum* to represent Papaveraceae and monocots, respectively, was highly significant ($\chi^2 = 27.5$, $P < 0.0001$) and placed the point of recombination midway in the gene (Fig. 3). Other genera in the Papaveraceae contain only a non-chimaeric, vertically transmitted *rps11* gene (Fig. 2d, e), making this transfer evolutionarily recent.

Finding four cases of HGT for just two mitochondrial genes, each only modestly sampled taxonomically, implies that HGT occurs at an appreciable frequency for plant mitochondrial genes in general. Indeed, perusal of the limited literature on plant mitochondrial phylogenies identified a fifth, strongly supported but previously misinterpreted, case of HGT. *Amborella trichopoda*, the sole extant member of the earliest branch of angiosperm evolution (see, for example, refs 18, 19), was reported by two different groups to contain a mitochondrial *atp1* gene of anomalous phylogenetic placement within eudicots. One study¹⁸ attributed this placement to the *Amborella* gene's 'divergent sequence', implying that it was misplaced as an artefact of long-branch attraction. The other study¹⁹ found an additional mitochondrial *atp1* gene in *Amborella*, of expected basal position; the authors invoked gene duplication, calling the eudicot-like *atp1* gene a 'paralogue' of this basal gene.

We independently isolated the same two *atp1* genes from *Amborella* as those reported in ref. 19, and our phylogenetic analyses strongly support a eudicot placement for one of these genes (Fig. 2f). We are convinced that neither published explanation is correct and

that instead the 'misplaced' *atp1* duplicate in *Amborella* is the result of HGT from eudicots. If long-branch attraction were the explanation, then this gene should itself represent a long branch and should group with other long-branched *atp1* genes. This does not happen (Fig. 2f). Moreover, that this gene is more similar to *atp1* genes of eudicots than to *atp1* genes of basal angiosperms also negates long-branch attraction. Paralogy as an explanation fails because, given the basal position of *Amborella* among angiosperms, this implies gene duplication in a common ancestor of angiosperms. If so, the two *Amborella* genes should each branch at the base of two separate clades of angiosperm *atp1* genes, each containing various diverse angiosperms (even allowing differential gene loss), in a pattern recapitulating the generally understood branching pattern of angiosperm phylogeny. This is clearly not so (Fig. 2f), leaving HGT as the only viable explanation.

We examined the expression status of two of the five cases of HGT. Fifteen *rps11* complementary DNAs were sequenced from *Sanguinaria* and each was found to be identical to its chimaeric *rps11* gene except for five sites of partial C → U RNA editing (Table 1). This result establishes that the chimaeric *Sanguinaria* gene is both transcribed and RNA edited, and indicates, in concert with its being an intact open reading frame, that it is probably functional. The horizontally acquired *atp1* gene of *Amborella* is also transcribed and RNA edited (U.B., C. Mathews, and J.D.P., unpublished observations). Roughly half of the genes characterized in the three other HGT cases are intact and therefore merit studies on their expression, whereas half show signs of being pseudogenes. The intact open reading frames include four of five Caprifoliaceae *rps11* genes and one of four Betulaceae *rps11* genes (the non-intact genes each contain a single frameshift mutation of four or five nucleotides). The phylogenetic mixture of both intact and probably

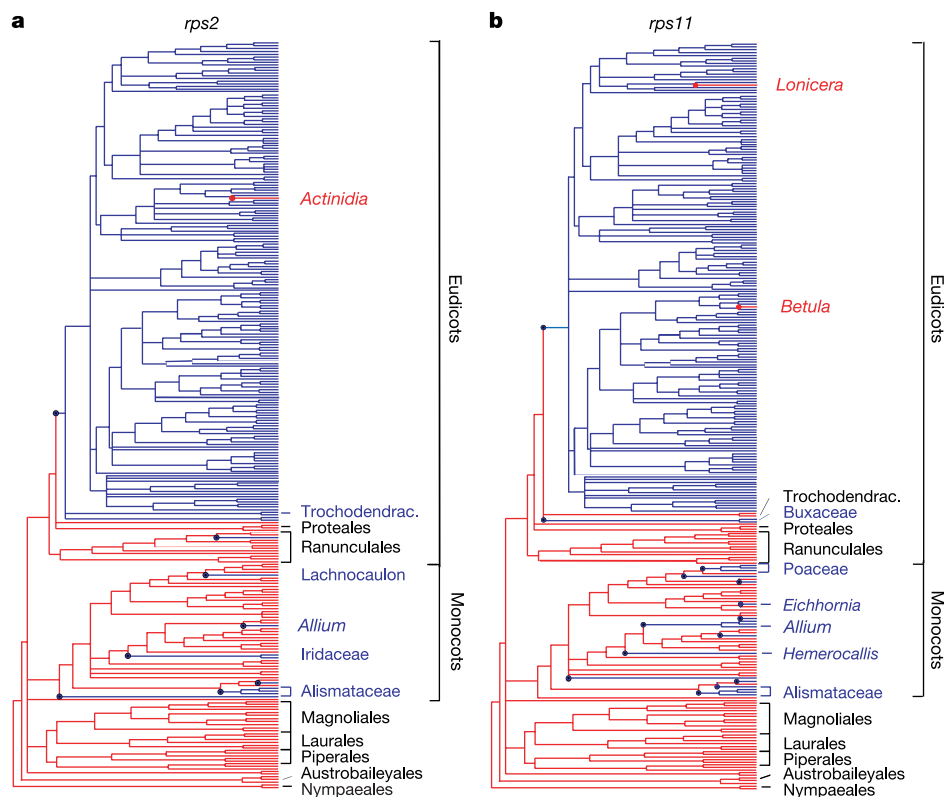


Figure 1 Anomalous presence of ribosomal protein genes in three angiosperm mitochondrial DNAs. A consensus phylogeny of 280 angiosperms is marked according to the presence (red branches) or absence (blue branches) of *rps2* (a) and *rps11* (b) in mitochondrial DNA (tree topology and gene presence/absence data are from ref. 13). Blue

and red bullets mark inferred losses and regains, respectively, of these genes. Names of taxa with gene regain are shown in red lettering, selected taxa with gene loss are in blue, and names of major groups of angiosperms are in black. Trochodendraceae.

disabled genes in these two families might mean that these are all non-functional genes, with only some having already been hit by disabling mutations. Alternatively, it might reflect differential usage and fixation of a duplicated gene (relative to a transferred nuclear homologue; see Supplementary Fig. 1), perhaps analogous to the situation reported for the transcompartmental *cox2* gene family created by gene transfer from mitochondrion to nucleus in legumes²⁰. Finally, all *Actinidia* *rps2* genes contain a single NT substitution early in the gene that creates a stop codon, unless remedied by rare U → C RNA editing.

Artefacts of DNA contamination or mislabelled samples, always a concern when invoking HGT, can be ruled out in all five transfer cases because multiple sampling (see Fig. 2b, for example) showed

Table 1 RNA editing of *Sanguinaria rps11*

Site (nucleotide)	Codon change	Efficiency*
78†	TTC (Phe) → TTT (Phe)	3/15
92†	TCG (Ser) → TTG (Leu)	8/15
143	CCA (Phe) → CTA (Leu)	9/15
146	CCG (Pro) → CTG (Leu)	8/15
351†	TTC (Phe) → TTT (Phe)	7/15

*Efficiency is shown as the fraction of the 15 cDNA clones that have been edited at a particular site; although editing efficiency at any one site did not exceed 60%, 7 of 15 cDNAs were edited at all three sites of non-synonymous editing.

†Sites that are also edited in rice *rps11* (ref. 30); the other two editing changes conserve the amino acids coded at these positions in monocots.

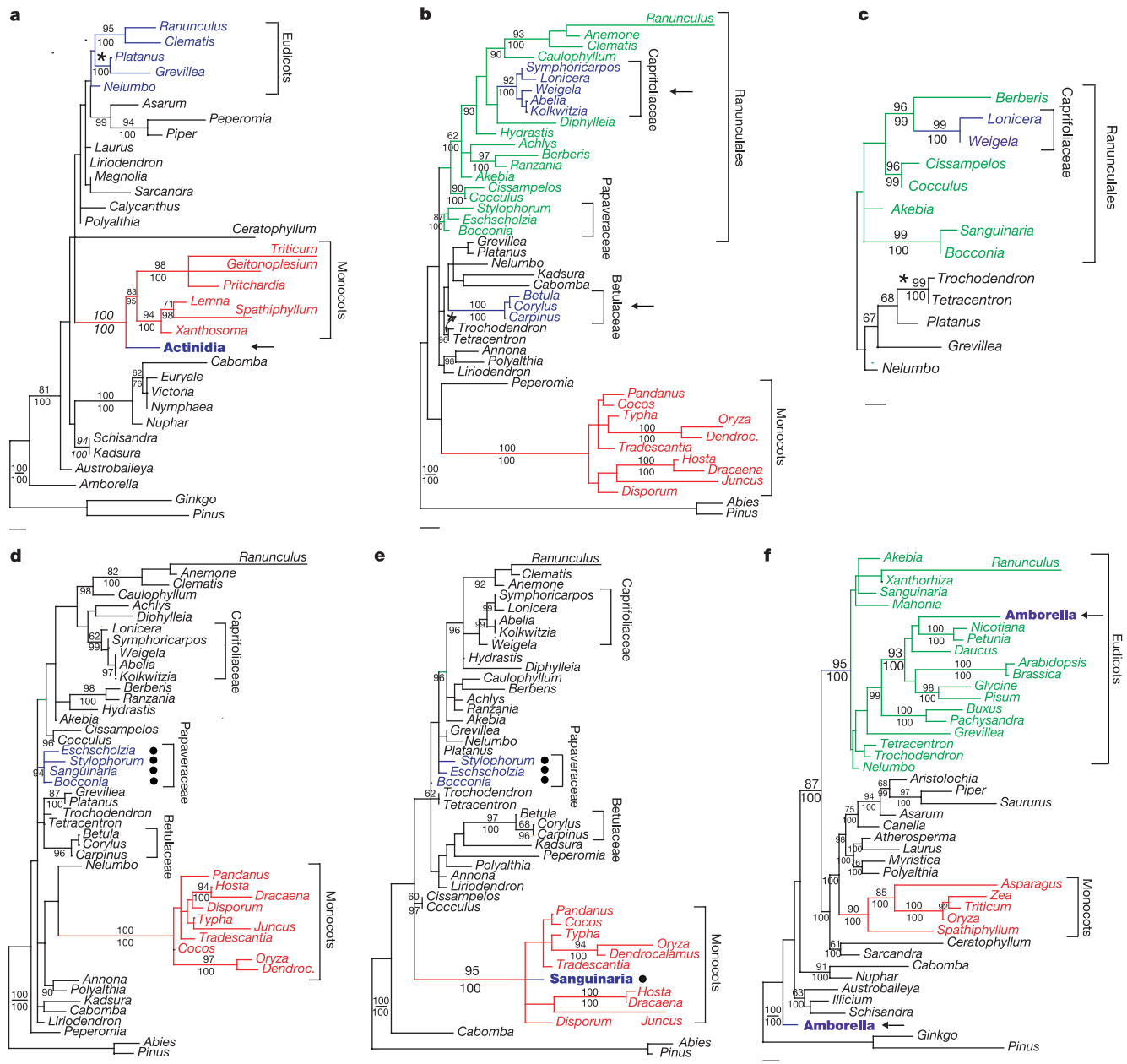


Figure 2 Phylogenetic evidence for HGT in angiosperm mitochondrial DNA. Maximum likelihood trees of *rps2* (474-nucleotide alignment) (a), *rps11* (456 nucleotides) (b), sequences immediately upstream of *rps11* (457 nucleotides) (c), 5' half of *rps11* (219 nucleotides) (d), 3' half of *rps11* (237 nucleotides) (e) and *atp1* (1,254 nucleotides) (f). *Dendroc.*, *Dendrocalamus*. Bootstrap support values more than 60% from parsimony

analyses are given above nodes, and bayesian posterior probability values more than 90% are given below. All scale bars correspond to 0.01 nucleotide substitutions per site. Asterisks in a–c indicate the positions of *Actinidia rps2* and Caprifoliaceae and Betulaceae *rps11* expected according to models of vertical transmission.

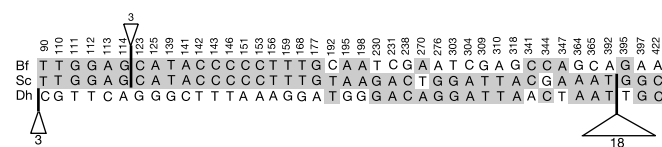


Figure 3 Chimaeric structure of the *Sanguinaria rps11* gene. Shown are all *rps11* variations in two Papaveraceae (*Bocconia frutescens* (Bf) and *Sanguinaria canadensis* (Sc)) and the monocot *Disporum hookeri* (Dh). Shading marks taxa that have the same nucleotide at a given position, numbered according to its location in *Sanguinaria*. Triangles mark deletions. The six variations (sites 192, 270, and so on) that do not follow the general shaded patterns are uniquely derived among either all 44 *rps11* genes sequenced (positions 192, 344 and 395, see Supplementary Fig. 3), or within the relevant group (Ranunculales for position 270, and monocots for positions 341 and 347; see Supplementary Fig. 3).

the results to be entirely reproducible (see Supplementary Information for details). Evidence that all five transferred genes are located in the mitochondrial genome and were horizontally acquired from mitochondrial rather than nuclear genomes relates to three factors: their lack of divergence (Fig. 2 and Supplementary Fig. 1), their hybridization intensity¹³ and their RNA editing (for at least *Sanguinaria rps11* and *Amborella atp1*), all of which are mitochondrion-like (see Supplementary Information for details).

Here we have identified strong evidence for four cases of plant-to-plant horizontal transfer of mitochondrial genes, and weaker evidence for a fifth. Three transfers involve the recapture of a gene lost early during eudicot evolution owing to functional transfer to the nucleus ('recapture HGT'), whereas *Amborella* contains intact genes of vertical and horizontal transmission ('duplicative HGT') and two such genes have recombined in *Sanguinaria* to create a strikingly chimaeric and expressed gene ('chimaeric HGT'). All five cases involve wide HGT within the context of angiosperm evolution (Fig. 4), including two transfers from monocots to eudicots. On the basis of current sampling and molecular-clock-based divergence

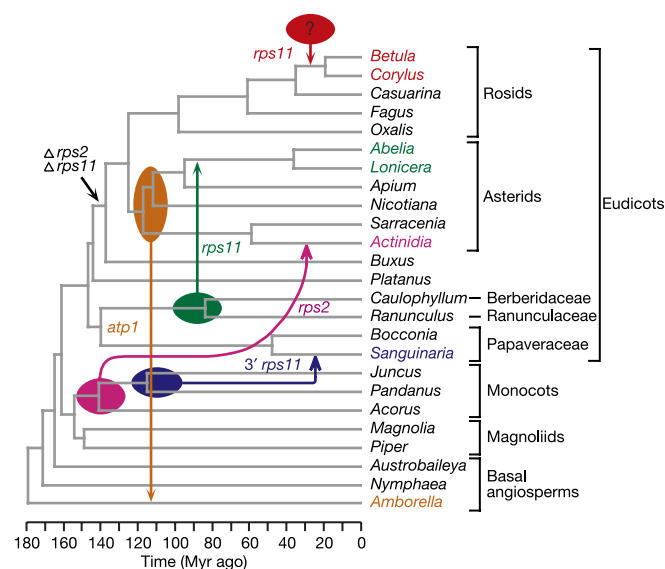


Figure 4 Approximate timing and donor-recipient relationships of five HGT 'events' in angiosperm mitochondrial DNA. Divergence times are from ref. 21. Shaded ovals indicate rough identity of donor groups (Fig. 2). The exact placement of arrowheads on recipient lineages is arbitrary. If correct, the older ages of donors relative to recipients for the *rps2* and 3' *rps11* transfers imply the existence of the transferred gene in an intermediate, unidentified vectoring agent or host plant for millions of years, but these discrepancies could easily be due to imprecision in the gene trees (Fig. 2) and/or in molecular-clock-based estimates²¹ of divergence times.

times²¹, we can very roughly estimate the age(s) of each transfer 'event' (Fig. 4). Further sampling promises to improve the precision of these estimates, to improve the identification of donor and recipient groups and to identify cases of long-term residency of a transferred gene in an intermediate genome, either of a vectoring agent or another plant group (if donors are found to be convincingly older than recipients for any well-dated cases; see Fig. 4 for two potential cases).

These results establish for the first time that conventional genes are subject to evolutionarily frequent HGT during plant evolution and provide the first unambiguous evidence that plants can donate DNA horizontally to other plants (compare refs 12 and 22 on both issues). This is also the best evidence (see also ref. 23) that eukaryotic genomes regularly acquire genes by means of horizontal events that are relatively recent, datable, and definable as to donor and recipient. For several reasons (see Supplementary Information) we believe the five cases reported here are merely the tip of a large iceberg of mitochondrial HGT in plants. Given this and the evolutionarily frequent occurrence of IGT to plant nuclear genomes^{13,16,24,25}, it seems likely that plant nuclear genomes are also significantly affected by HGT. Indeed, a few cases of horizontal acquisition of bacterial genes by plant nuclear genomes have been reported^{23,26}. Despite extensive phylogenetic analysis of chloroplast genes, there is no published evidence for the acquisition of foreign DNA by chloroplasts in any land plant. We therefore predict a much lower incidence of chloroplast HGT than for mitochondrial or nuclear genomes. It is fortunate that the two major sets of genes used to reconstruct plant phylogeny—chloroplast genes and nuclear rRNA genes—seem relatively immune to HGT.

Our findings raise many other questions. Are these results relevant to concerns over the potential escape of transgenes from genetically modified plants by means of HGT? We think not, because although reasonably frequent on an evolutionary time scale of millions of years, HGT is highly unlikely to be a factor on a human time scale. Are certain plants especially susceptible to HGT, as is clearly true for IGT¹³? Does HGT ever occur on a grand scale, leading to the horizontal acquisition of much or all of a mitochondrial genome, and/or of many nuclear genes, as has been seen for IGT²⁵? How do genes move from one plant to another, sexually unrelated, plant? Is HGT driven predominantly by potential vectoring agents such as viruses, bacteria, fungi, insects, pollen or even meteorites; or by the transformational uptake of plant DNA released into the soil; or by unrelated plants occasionally grafting together? □

Methods

Gene isolation and characterization

Mitochondrial *rps2*, *rps11* and *atp1* genes were amplified by standard, direct polymerase chain reaction (PCR) in an Idaho Technologies Air Thermocycler. In general, each reaction consisted of 36 cycles of 10 s at 94 °C, 15 s at 50 °C and 45 s at 72 °C. The extension time was 90 s and the annealing temperature was 55 °C for *atp1* amplification and for inverse PCR. A list of PCR primers can be found in Supplementary Information. Sequences flanking *rps2* from *Actinidia arguta* were obtained by inverse PCR: *A. arguta* DNA (2 µg) was digested with either *ApoI* or *BamHI* plus *BglI* (New England Biolabs) and then ligated overnight at 12 °C with T4 ligase (New England Biolabs) in 400 µl. After extraction with chloroform and precipitation with ethanol, the ligated DNA was used for inverse PCR. Primers to conserved sequences immediately upstream of *rps11* in *Lonicera* and *Sanguinaria* (initially amplified by Vectorette PCR; Sigma-Genosys) were used to amplify this region in other species. To verify the authenticity of the *Sanguinaria rps11* sequence, PCR was performed on DNA isolated from three independent sources, and one set of *Sanguinaria* DNA extractions and PCR reactions were done in a different laboratory. *Sanguinaria* RNA was isolated from roots and flower buds using RNeasy Plant Mini Kit (Qiagen) and treated twice with DNase I (TaKaRa). Reverse transcription was done on 2 µg RNA with random hexamer primers (Invitrogen) and Moloney-murine-leukaemia virus reverse transcriptase (New England Biolabs). Gel-purified RT-PCR products (Qiaquick; Qiagen) of *Sanguinaria rps11* were cloned with TOPO TA Cloning (Invitrogen) in accordance with the manufacturer's directions. Uncloned PCR products and cloned RT-PCR products were sequenced on both strands at the DNA sequencing facility of the Indiana Molecular Biology Institute.

Phylogenetic analyses

Phylogenetic analyses were performed with PAUP v. 4.0b10 (ref. 27) and MrBayes v. 2.01 (ref. 28). Sites of known RNA editing were excluded from the analyses. Maximum-likelihood trees were constructed with a HKY85 substitution model; trees in Fig. 2 used a transition-to-transversion ratio of 2.0. For alternative topology tests (see below for Shimodaira–Hasegawa tests, and Supplementary Information for parametric bootstrap analyses), transition-to-transversion ratios and gamma distribution parameters were first estimated from the data. Bootstrap support values are from maximum-parsimony analyses of 1,000 bootstrap replicates and 100 random addition replicates. Posterior probability for clade support was estimated with Markov chain Monte Carlo as implemented in MrBayes. Four Markov chains were run for 10^5 to 10^6 generations after burn-in, using random initial trees and a general time-reversible (GTR) codon-site-specific substitution model for coding sequences and GTR with gamma distribution for non-coding sequences. The Shimodaira–Hasegawa²⁹ test favoured the horizontal placements shown in the unconstrained trees of Fig. 2 over alternative topologies based on vertical transmission: first, the unconstrained *rps2* tree (Fig. 2a) was favoured over the constrained tree grouping *Actinidia* with *Grevillea*/*Platanus* as a monophyletic group ($P = 0.012$); second, the unconstrained *rps11* tree (Fig. 2b) was favoured over the tree grouping Caprifoliaceae plus Betulaceae plus Trochodendraceae and/or Proteales ($P = 0.036$); third, the unconstrained 3' *rps11* tree (Fig. 2e) was favoured over the tree grouping *Sanguinaria* with other Papaveraceae ($P < 0.001$); and last, the unconstrained *atp1* tree (Fig. 2f) was favoured over the tree grouping the two *Amborella* sequences ($P < 0.001$). A test of recombination was performed with Maximum Chi-squared for Macintosh, version 1.0, by N. Ross, which implements the original method by J. Maynard Smith¹⁷.

Received 27 February; accepted 13 May 2003; doi:10.1038/nature01743.

- Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).
- Koonin, E. V., Makarova, K. S. & Aravind, L. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu. Rev. Microbiol.* **55**, 709–742 (2001).
- Nixon, J. E. *et al.* Evidence for lateral transfer of genes encoding ferredoxins, nitro-reductases, NADH oxidase, and alcohol dehydrogenase 3 from anaerobic prokaryotes to *Giardia lamblia* and *Entamoeba histolytica*. *Eukaryotic Cell* **1**, 181–190 (2002).
- Andersson, J. O., Sjögren, Å. M., Davis, L. A. M., Embley, T. M. & Roger, A. J. Phylogenetic analyses of diplomonad genes reveal frequent lateral gene transfers affecting eukaryotes. *Curr. Biol.* **13**, 94–104 (2003).
- Archibald, J. M., Rogers, M. B., Toop, M., Ishida, K.-I. & Keeling, P. J. Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga, *Bigelowiella natans*. *Proc. Natl Acad. Sci. USA* **100**, 7678–7683 (2003).
- Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
- Stanhope, M. J. *et al.* Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature* **411**, 940–944 (2001).
- Salzberg, S. L., White, O., Peterson, J. & Eisen, J. A. Microbial genes in the human genome: lateral transfer or gene loss? *Science* **292**, 1903–1906 (2001).
- Robertson, H. M. in *Mobile DNA II* (eds Craig, N. L., Craigie, R., Gellert, M. & Lambowitz, A. M.) 1093–1110 (ASM, Washington DC, 2002).
- Eickbush, T. H. & Malik, H. S. in *Mobile DNA II* (eds Craig, N. L., Craigie, R., Gellert, M. & Lambowitz, A. M.) 1111–1144 (ASM, Washington DC, 2002).
- Goddard, M. R. & Burt, A. Recurrent invasion and extinction of a selfish gene. *Proc. Natl Acad. Sci. USA* **96**, 13880–13885 (1999).
- Cho, Y., Qiu, Y.-L., Kuhlman, P. & Palmer, J. D. Explosive invasion of plant mitochondria by a group I intron. *Proc. Natl Acad. Sci. USA* **95**, 14244–14249 (1998).
- Adams, K. L., Qiu, Y. L., Stoutemyer, M. & Palmer, J. D. Punctuated evolution of mitochondrial gene content: High and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. *Proc. Natl Acad. Sci. USA* **99**, 9905–9912 (2002).
- Wolfe, K. H., Li, W. S. & Sharp, P. M. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc. Natl Acad. Sci. USA* **84**, 9054–9058 (1987).
- Laroche, J., Li, P., Maggia, L. & Bousquet, J. Molecular evolution of angiosperm mitochondrial introns and exons. *Proc. Natl Acad. Sci. USA* **94**, 5722–5727 (1997).
- Adams, K. L., Daley, D. O., Qiu, Y. L., Whelan, J. & Palmer, J. D. Repeated, recent and diverse transfers of a mitochondrial gene to the nucleus in flowering plants. *Nature* **408**, 354–357 (2000).
- Maynard Smith, J. Analyzing the mosaic structure of genes. *J. Mol. Evol.* **34**, 126–129 (1992).
- Qiu, Y. L. *et al.* The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature* **402**, 404–407 (1999).
- Barkman, T. J. *et al.* Independent and combined analyses of sequences from all three genomic compartments converge on the root of flowering plant phylogeny. *Proc. Natl Acad. Sci. USA* **97**, 13166–13171 (2000).
- Adams, K. L. *et al.* Intracellular gene transfer in action: Dual transcription and multiple silencings of nuclear and mitochondrial *cox2* genes in legumes. *Proc. Natl Acad. Sci. USA* **96**, 13863–13868 (1999).
- Wikstrom, N., Savolainen, V. & Chase, M. W. Evolution of the angiosperms: Calibrating the family tree. *Proc. R. Soc. Lond. B* **268**, 2211–2220 (2001).
- Feschotte, C. & Wessler, S. R. Mariner-like transposases are widespread and diverse in flowering plants. *Proc. Natl Acad. Sci. USA* **99**, 280–285 (2002).
- Intrieri, M. C. & Buiatti, M. The horizontal transfer of *Agrobacterium rhizogenes* genes and the evolution of the genus *Nicotiana*. *Mol. Phylogenet. Evol.* **20**, 100–110 (2001).
- Blanchard, J. L. & Schmidt, G. W. Pervasive migration of organellar DNA to the nucleus in plants. *J. Mol. Evol.* **41**, 397–406 (1995).
- Stupar, R. M. *et al.* Complex mtDNA constitutes an approximate 620-kb insertion on *Arabidopsis thaliana* chromosome 2: Implication of potential sequencing errors caused by large-unit repeats. *Proc. Natl Acad. Sci. USA* **98**, 5099–5103 (2001).
- Zardoya, R., Ding, X. D., Kitagawa, Y. & Chrispeels, M. J. Origin of plant glycerol transporters by horizontal gene transfer and functional recruitment. *Proc. Natl Acad. Sci. USA* **99**, 14893–14896 (2002).

- Swofford, D. L. PAUP*: Phylogenetic analysis using parsimony (*and other methods) Version 4 (Sinauer, Sunderland, Massachusetts, 1998).
- Huelsenbeck, J. P. & Ronquist, F. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755 (2001).
- Shimodaira, H. & Hasegawa, M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**, 1114–1116 (1999).
- Kadowaki, K., Kubo, N., Ozawa, K. & Hirai, A. Targeting presequence acquisition after mitochondrial gene transfer to the nucleus occurs by duplication of existing targeting signals. *EMBO J.* **15**, 6652–6666 (1996).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank C. Mathews for technical assistance; E. Knox for creating Fig. 4, for drafting Fig. 1 and for discussion; L. Goertzen, E. Knox, R. Olmstead, D. Rice and S. Stefanovic for critical reading of the manuscript; R. Gardner for providing several *Actinidia* DNAs; M. Stoutemyer and J. Gastony for help in obtaining plant material; and B. Hall and the University of California Santa Cruz arboretum for supplying leaf material for *Amborella*. Financial support was provided by the US NIH.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.D.P. (jpalmer@bio.indiana.edu). The sequences reported in this study have been deposited in GenBank under the accession numbers AY293002–AY293057.

Topography and synaptic shaping of direction selectivity in primary auditory cortex

Li I. Zhang^{*†}, Andrew Y. Y. Tan^{*†}, Christoph E. Schreiner^{*} & Michael M. Merzenich^{*}

^{*} Coleman Memorial Laboratory and W.M. Keck Foundation Center for Integrative Neuroscience, University of California, San Francisco, California 94143, USA

[†] These authors contributed equally to this work

The direction of frequency-modulated (FM) sweeps is an important temporal cue in animal and human communication. FM direction-selective neurons are found in the primary auditory cortex (A1)^{1,2}, but their topography and the mechanisms underlying their selectivity remain largely unknown. Here we report that in the rat A1, direction selectivity is topographically ordered in parallel with characteristic frequency (CF): low CF neurons preferred upward sweeps, whereas high CF neurons preferred downward sweeps. The asymmetry of ‘inhibitory sidebands’, suppressive regions flanking the tonal receptive field (TRF) of the spike response, also co-varied with CF. *In vivo* whole-cell recordings showed that the direction selectivity already present in the synaptic inputs was enhanced by cortical synaptic inhibition, which suppressed the synaptic excitation of the non-preferred direction more than that of the preferred. The excitatory and inhibitory synaptic TRFs had identical spectral tuning, but with inhibition delayed relative to excitation. The spectral asymmetry of the synaptic TRFs co-varied with CF, as had direction selectivity and sideband asymmetry, and thus suggested a synaptic mechanism for the shaping of FM direction selectivity and its topographic ordering.

Extracellular multiunit spike responses to sweeps of various speeds and intensities were recorded in the mid-layers of the adult rat A1. Responses from a representative low CF site are shown in Fig. 1. Sweeps of different speeds evoked distinct responses (Fig. 1a). The onset and duration of each response mostly reflected the timing of the sweep’s intersection with the TRF of the spike response (Fig. 1b). A direction selectivity index (DSI) was calculated for

responses to pairs of up and down sweeps of the same speed: a positive DSI indicates up-sweep selectivity (see Methods). The strongest direction selectivity was obtained for 70 octaves s^{-1} sweeps, regardless of sweep intensity (Fig. 1c)³. Accordingly, we chose sweeps of speed 70 octaves s^{-1} and intensity 60 dB to further characterize direction selectivity in A1.

The primary auditory cortex was mapped. The CFs of neurons increased smoothly along the posterior–anterior axis, with low CFs at the posterior end (Fig. 2a)^{4,5}. Notably, FM direction selectivity was topographically ordered in parallel with CF. DSI changed continuously and monotonically along the CF axis (Fig. 2b, c). Low CF sites (<8 kHz) responded more strongly to up sweeps, whereas high CF sites (>14 kHz) responded more strongly to down sweeps.

Any mechanism that generates FM direction selectivity must possess both spectral and temporal asymmetries. In bats and ferrets, spectrally asymmetric inhibition co-varies with direction selectivity^{6–8}. We examined whether this was also the case in the rat A1 by determining how the frequency and intensity of one tone affected its ability to suppress responses to a subsequent second tone^{6–10}. Figure 3a shows the inhibitory areas (outlined by blue) obtained for a low CF site and a high CF site. For the low CF site, the inhibitory region covered the whole TRF (in red) and extended beyond the high frequency boundary of the TRF; for the high CF site, the inhibitory region had a greater extension on the low frequency side of the TRF. The suppressive regions flanking the TRF are known as ‘inhibitory sidebands’. As low CF neurons prefer up sweeps and high CF neurons down sweeps, these two particular neurons are consonant with the notion that inhibitory sideband asymmetry is

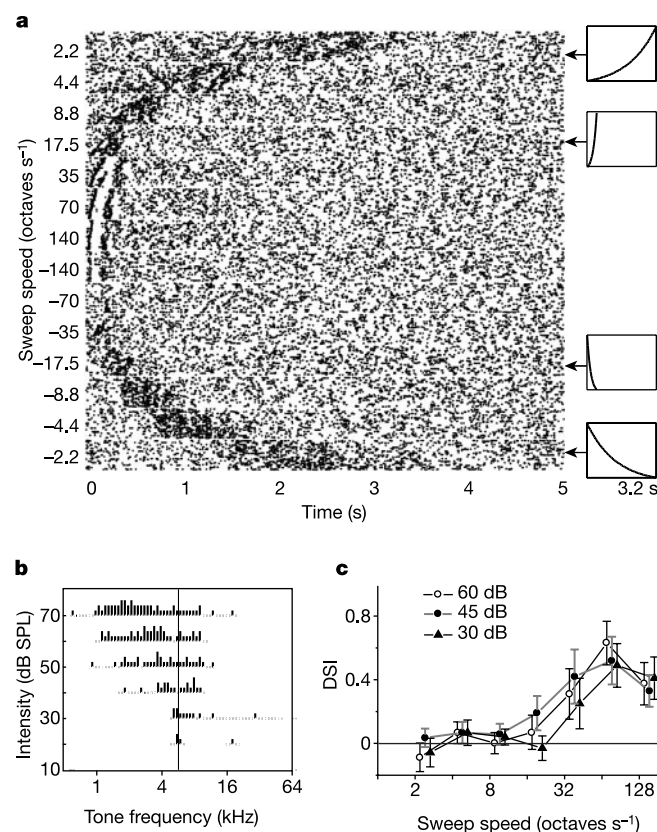


Figure 1 Direction selectivity of multiunit sites. **a**, Rasters of responses from one site to 7-octave FM sweeps. The sweep speed of each panel is on the left. Positive speeds indicate up sweeps. Within each panel are three repetitions of 15 intensities evenly spaced from 30 (top) to 75 dB (bottom). Schematic frequency versus time representations of four sweeps are on the right. **b**, TRF for the same site. Bar lengths indicate the number of spikes evoked by each tone. The vertical line indicates the CF. SPL, sound pressure level. **c**, DSI versus sweep speed for the same site at different sweep intensities.

correlated with direction selectivity. To assess the CF dependence of inhibitory sideband asymmetry, we measured the bandwidths of the lower and upper inhibitory sidebands at an intensity of 60 dB. Figure 3b shows the data for 84 sites grouped in 1-octave bins. The width of the lower inhibitory sideband increased with CF, whereas the width of the upper inhibitory sideband decreased with CF. Inhibitory sideband asymmetry therefore co-varies with CF-dependent direction selectivity (Fig. 2d). This strongly suggests that the spectral asymmetry of inhibition is a source of the spectral asymmetry required for direction selectivity.

Inhibitory sidebands arise from inhibition accumulated along the central pathways leading up to and including A1, and could be due to inhibition from both synaptic inputs and intrinsic conductances^{9,10}. To explore the contribution of cortical synapses to CF-dependent direction selectivity, *in vivo* whole-cell voltage-clamp recordings were carried out in A1. Figure 4a shows the synaptic currents evoked by 70 octaves s^{-1} sweeps for a low CF neuron. Excitatory inputs were examined by clamping the neuron at

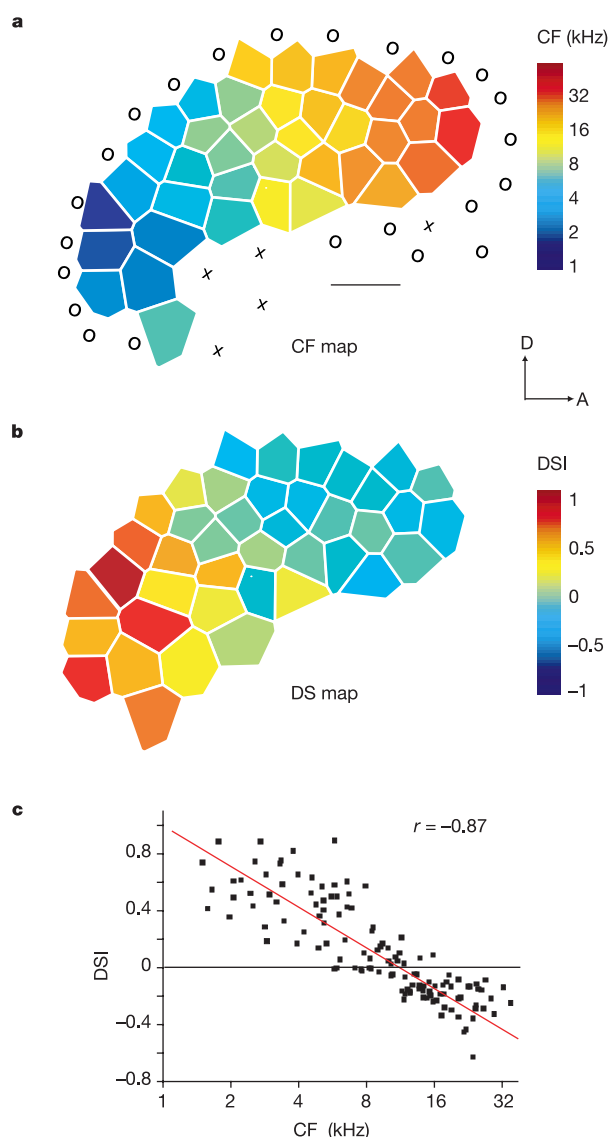


Figure 2 CF dependence of direction selectivity in A1. **a**, CF map. Colours indicate the CFs of cortical locations. Sites without responses to tones or clear tuning are represented by ‘o’ and ‘x’ respectively. Scale bar, 0.5 mm. A, anterior; D, dorsal. **b**, Direction selectivity (DS) map from the same cortex. Colours indicate the DSI. **c**, DSI CF dependence of 123 sites from five rats. Least squares fit, with correlation coefficient r , is shown in red.

–70 mV (upper panels), whereas inhibitory inputs were revealed as outward-going currents at –30 mV (lower panels). Both excitatory and inhibitory inputs were direction selective, and had the same direction preference. The DSI of the excitatory current for the neuron of Fig. 4a was greatest for sweep speed of 70 octaves s^{-1} (Fig. 4b). Across nine neurons the DSIs of the excitatory current for 70 octaves s^{-1} sweeps decreased as a function of CF (Fig. 4c).

The direction selectivity of the synaptic excitation presumably reflects direction selectivity that is already present at subcortical stations^{11–13}. However, the extent to which synaptic excitation is thalamic or cortical in origin cannot be inferred from our data. Synaptic inhibition is, on the other hand, purely cortical in origin. As cortical synaptic inhibition may, together with synaptic excitation, influence the spike response, we examined the interaction of synaptic excitation with cortical synaptic inhibition.

Excitatory and inhibitory conductances were derived from the traces of Fig. 4a (see Methods). The linear current–voltage (I – V) curve of Fig. 4d demonstrates that the neuron was reasonably well clamped. A 70 octaves s^{-1} sweep in the preferred direction produced a fast, large and transient increase in the excitatory conductance, which was followed by a large inhibitory conductance (Fig. 4e, left panels). The non-preferred direction evoked less excitation and inhibition (Fig. 4e, right panels). For the preferred direction, the inhibitory peak occurred well after the excitatory peak. For the non-preferred direction, there was, interestingly, a greater overlap of the excitation by inhibition, with the inhibitory peak closer in time to the excitatory peak; in this particular neuron, the two peaks were in fact practically co-incident. A similar difference between the two directions was exhibited by the seven neurons of Fig. 4c that had strong direction selectivity (Fig. 4f; paired t -test, $P < 0.01$). Therefore, by suppressing the synaptic excitation of the non-preferred direction more than that of the

preferred, cortical synaptic inhibition can enhance direction selectivity.

To characterize further how the interaction of excitation and inhibition enhances direction selectivity, we examined the synaptic TRFs, as we expected them to be major determinants of the sweep responses^{14–16}. Figure 5a, b shows the synaptic currents evoked by pure tones of various frequencies and intensities for a low and high CF neuron, respectively. An inhibitory current was always preceded by an excitatory current, with strong inhibition following strong excitation. The strengths of the excitatory and inhibitory currents were strongly correlated (Fig. 5c). In all eight A1 neurons for which complete synaptic TRFs were obtained, the TRFs of both excitatory and inhibitory currents had nearly identical spectral tuning, consistent with previous results^{17–19}.

In addition, the TRFs of excitatory and inhibitory currents exhibited the same spectral asymmetry. For the low CF neuron, strong currents clustered at low frequencies, with a long tail of weaker responses at higher frequencies (Fig. 5a). The asymmetry was reversed for the high CF neuron, with strong currents clustered at high frequencies (Fig. 5b). We characterized asymmetries in the synaptic TRFs by the skewness of the iso-intensity profile at 60 dB (see Methods). A positive skewness indicates a long tail of small responses extending into high frequencies. For 18 neurons the skewness of the synaptic TRF changed monotonically with CF, changing sign at approximately 12 kHz (Fig. 5d). Synaptic TRF

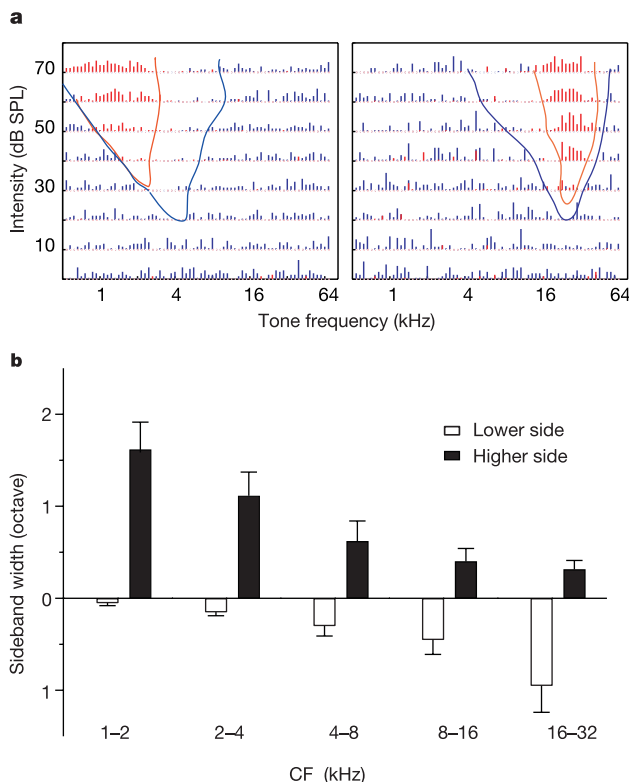


Figure 3 CF dependence of inhibitory sideband asymmetry. **a**, The left panel shows a low CF neuron. Inhibitory areas are outlined by blue. Blue bars indicate responses to the second tone as a function of the frequency and intensity of the first. Red bars show the TRF. **b**, Mean width of inhibitory sidebands as a function of CF.

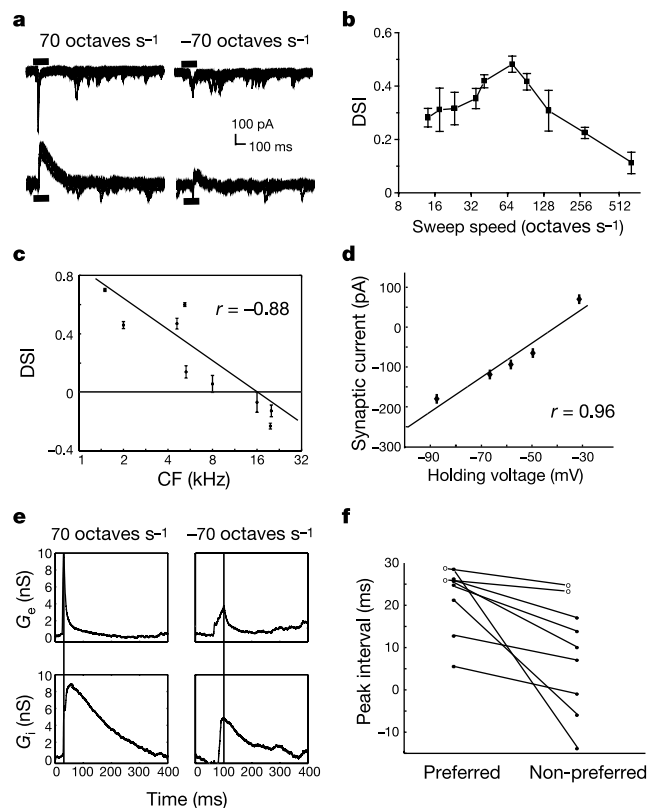


Figure 4 Synaptic responses to FM sweeps. **a**, Synaptic currents at –70 mV (upper panels) and –30 mV (lower panels) evoked by up and down sweeps (left and right panels respectively) in a 2-kHz neuron. Five stimulus repetitions are superimposed. **b**, DSI of excitatory current versus sweep speed. **c**, DSI of excitatory current versus CF. **d**, I – V curve at 20 ms after up-sweep onset. **e**, Excitatory (upper panels) and inhibitory (lower panels) conductances evoked by up and down sweeps. Vertical lines indicate peaks of excitatory conductances. **f**, Time by which the inhibitory peak follows the excitatory peak for the preferred (left) and non-preferred (right) directions for strongly (filled circles) or mid-CF, weakly (open circles) direction-selective neurons. Points for the same neuron are connected.

asymmetry therefore co-varies with CF-dependent direction selectivity (Fig. 2d). This indicates that the spectral asymmetry of synaptic TRFs is responsible for some of the spectral asymmetry required for direction selectivity.

In fact, the spectral asymmetry of the synaptic TRFs, combined with the temporal asymmetry provided by the delay of inhibition relative to excitation, suggests a mechanism for direction selectivity. Consider a model low CF neuron with synaptic TRF asymmetry: let it have strong inputs on the low frequency side, and a long tail of weak inputs on the high frequency side. An up sweep first evokes strong excitation that could trigger spiking before the onset of the following strong inhibition. But a down sweep will first evoke only weak excitation; the following weak inhibition will extend into the onset of the strong excitation, reducing its ability to cause spiking. Thus, this model generates direction selectivity that is consistent with the observed co-variation of direction selectivity and synaptic TRF asymmetry (Fig. 5d). It does so with inhibition suppressing the excitation of the non-preferred direction more than that of the preferred, consistent with how direction selectivity is enhanced by A1 neurons (Fig. 4e, f).

The model also generates sideband asymmetry that is consistent with the observed co-variation of direction selectivity and sideband asymmetry (Fig. 3b). Sidebands can appear even if synaptic inhibition is not spectrally broader than synaptic excitation, because

synaptic inhibition contributes to inhibitory areas outside of the spike TRF if the excitation preceding the inhibition is subthreshold. By suggesting mechanisms for both direction selectivity and sideband asymmetry, synaptic TRF asymmetry unifies our previous observations and arguments (see also Supplementary Information).

The model, however, does not predict some features of the sweep responses, such as the direction selectivity of the synaptic conductances, or the duration of the excitatory conductance evoked by the preferred direction. This is presumably because it disregards both direction selectivity that is already present at subcortical stations^{11–13}, as well as nonlinearities of the cortical circuitry arising, for example, from recurrent connections. The excitatory synaptic input could be provided by an asymmetric convergence of frequency-tuned, monosynaptic, excitatory thalamic input onto the cortical neuron. The temporally delayed inhibition would be provided by the same thalamic input, di-synaptically relayed by local inhibitory neurons.

The asymmetric receptive fields, reminiscent of those observed in hippocampal place cells and visual tectal neurons after directional stimulation, might have emerged from spike-timing-dependent plasticity and directionally biased input^{20,21}. Note that asymmetric receptive fields can be generated in a recurrent circuit by spike-timing-dependent plasticity without any input bias²². The circuit amplifies small asymmetries in the initial conditions, which might be provided by boundary effects arising from the finite frequency range of the auditory system.

Whereas a topographical order of CF, retinal position, ocular dominance or orientation represents spectral or spatial information, that of direction selectivity represents inherently temporal information. Our data therefore shed light on the synaptic mechanisms underlying both the coding of temporal information by single neurons, as well as its conversion into large-scale spatial organization. □

Methods

Extracellular recording

All experimental procedures used in this study were approved under UCSF Animal Care Facility protocols. Experiments were carried out in a sound-attenuating chamber. Female Sprague-Dawley rats about 3 months old and weighing 280–300 g were anaesthetized with pentobarbital, and their right auditory cortex exposed. Multiunit spike responses were recorded with parylene-coated tungsten microelectrodes at 500–600 μm below the pial surface²³. Pure tones (0.5–64 kHz at 0.1-octave intervals, 25-ms duration) at eight 10-dB-spaced sound intensities were delivered through a calibrated free-field speaker facing the left ear. The frequencies used covered the hearing range of the rat.

FM sweep and two-tone stimuli

Logarithmic FM stimuli sweeping between 0.5–64 kHz with speeds of 2–700 octaves s^{-1} were generated and calibrated according to their output envelope. Sweeps were presented in a pseudorandom order. Except for the experiments in Fig. 1, sweep intensity was set at 60 dB. For two-tone stimuli, the first tone was varied as for determining the TRF. The second tone was a CF tone with intensity 5–10 dB above threshold, which could consistently trigger spiking responses, and was played 20 ms after the onset of the first. The method of measuring sideband widths was similar to those previously described^{8,9}. All sidebands lay entirely within the probed frequency range.

Whole-cell recording

A1 was first located by roughly mapping with a tungsten electrode. Whole-cell recordings^{24–27} were obtained from neurons located 400–700 μm beneath the cortical surface. We prevented cortical pulsation with 4% agarose. For voltage-clamp recording, the pipette (about 7 M Ω) contained (in mM): 125 Cs-glucuronate, 5 TeACl, 4 MgATP, 0.3 GTP, 10 phosphocreatine, 10 HEPES, 0.5 EGTA, 3.5 QX-314, 2 CsCl, pH 7.2. Recordings were made with an Axopatch 200B. The whole-cell capacitance was compensated and the initial series resistance (35–60 M Ω) was compensated to achieve effective series resistances of 15–30 M Ω . Signals were filtered at 5 kHz and sampled at 10 kHz. The CFs of the synaptic TRFs matched their positions in the rough CF map.

Data analysis

The DSI is defined as $(r_1 - r_2)/(r_1 + r_2)$, where r_1 is the response triggered by the up sweep and r_2 is the response to the down sweep.

The excitatory and inhibitory conductances were derived^{28–30} using $I(t) = G_e(V - E_e) + G_i(V - E_i)$. $I(t)$ is the amplitude of the synaptic current at time t , G_e and G_i are the resting conductance and membrane potential, $G_e(t)$ and $G_i(t)$ are the excitatory and inhibitory conductances, V is the holding voltage, and E_e (0 mV) and E_i

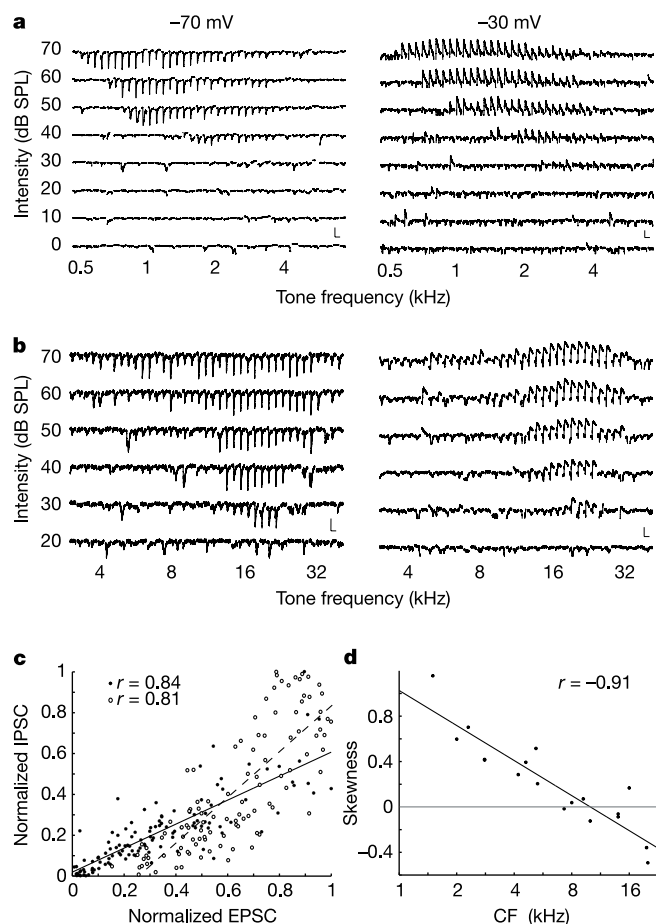


Figure 5 Asymmetric synaptic TRFs. **a**, Currents evoked in a low CF neuron at -70 (left panel) and -30 mV (right panel) to pure tones of various frequencies and intensities. Scale bars: 200 pA (vertical), 80 ms (horizontal) left; 100 pA (vertical), 400 ms (horizontal) right. **b**, Same as in **a**, but for a high CF neuron. Scale bars: 50 pA (vertical), 110 ms (horizontal) left; 40 pA (vertical), 400 ms (horizontal) right. **c**, Correlation of excitatory postsynaptic current (EPSC) and inhibitory postsynaptic current (IPSC) amplitudes. Filled circles are from recordings shown in **a**, open circles from **b**. **d**, Skewness versus CF.

(−65 mV) are the reversal potentials for excitatory and inhibitory currents. Changing E_e and E_i by ± 10 mV did not change the time course of any conductance significantly.

Skewness, a measure of the asymmetry based on the third central moment of a distribution, is defined as $m_3/m_2^{3/2}$, where $m_1 = \Sigma(r_f f)$, $m_2 = \Sigma(r_f(f - m_1)^2)$, $m_3 = \Sigma(r_f(f - m_1)^3)$ and $r_f = R_f/\Sigma(R_f)$. R_f is the synaptic current at −70 mV, averaged between 15 and 35 ms after the onset of a tone of frequency f , with f in units of octaves above 0.5 kHz. The sums are over all frequencies in the synaptic TRF at 60 dB. Skewness was calculated only for neurons whose synaptic TRFs lay entirely within the probed frequency range.

Received 10 February; accepted 7 May 2003; doi:10.1038/nature01796.

- Mendelson, J. R. & Cynader, M. S. Sensitivity of cat primary auditory cortex (AI) neurons to the direction and rate of frequency modulation. *Brain Res.* **327**, 331–335 (1985).
- Schreiner, C. E., Read, H. L. & Sutter, M. L. Modular organization of frequency integration in primary auditory cortex. *Annu. Rev. Neurosci.* **23**, 501–529 (2000).
- Ricketts, C., Mendelson, J. R., Anand, B. & English, R. Responses to time-varying stimuli in rat auditory cortex. *Hearing Res.* **123**, 27–30 (1998).
- Merzenich, M. M., Knight, P. L. & Roth, G. L. Representation of cochlea within primary auditory cortex in the cat. *J. Neurophysiol.* **38**, 231–249 (1975).
- Sally, S. L. & Kelly, J. B. Organization of auditory cortex in the albino rat: sound frequency. *J. Neurophysiol.* **59**, 1627–1638 (1988).
- Suga, N. Functional properties of auditory neurones in the cortex of echo-locating bats. *J. Physiol.* **181**, 671–700 (1965).
- Shamma, S. A., Fleschman, J. W., Wiser, P. R. & Versnel, H. Organization of response areas in ferret primary auditory cortex. *J. Neurophysiol.* **69**, 367–383 (1993).
- Nelken, I. & Versnel, H. Responses to linear and logarithmic frequency-modulated sweeps in ferret primary auditory cortex. *Eur. J. Neurosci.* **12**, 549–562 (2000).
- Calford, M. B. & Semple, M. N. Monaural inhibition in cat auditory cortex. *J. Neurophysiol.* **73**, 1876–1891 (1995).
- Brosch, M. & Schreiner, C. E. Time course of forward masking tuning curves in cat primary auditory cortex. *J. Neurophysiol.* **77**, 923–943 (1997).
- Clopton, B. M. & Winfield, J. A. Unit responses in the inferior colliculus of rat to temporal auditory patterns of tone sweeps and noise bursts. *Exp. Neurol.* **42**, 532–540 (1974).
- Britt, R. & Starr, A. Synaptic events and discharge patterns of cochlear nucleus cells. II. Frequency-modulated tones. *J. Neurophysiol.* **39**, 179–194 (1976).
- Poon, P. W. & Yu, P. P. Spectro-temporal receptive fields of midbrain auditory neurons in the rat obtained with frequency modulated stimulation. *Neurosci. Lett.* **289**, 9–12 (2000).
- Kowalski, N., Depireux, D. A. & Shamma, S. A. Analysis of dynamic spectra in ferret primary auditory cortex. II. Prediction of unit responses to arbitrary dynamic spectra. *J. Neurophysiol.* **76**, 3524–3534 (1996).
- deCharms, R. C., Blake, D. T. & Merzenich, M. M. Optimizing sound features for cortical neurons. *Science* **280**, 1439–1443 (1998).
- Schnupp, J. W., Morsic-Flogel, T. D. & King, A. J. Linear processing of spatial cues in primary auditory cortex. *Nature* **414**, 200–204 (2001).
- De Ribaupierre, F., Goldstein, M. H. Jr & Yeni-Komshian, G. Intracellular study of the cat's primary auditory cortex. *Brain Res.* **48**, 185–204 (1972).
- Volkov, I. O. & Galazjuk, A. V. Formation of spike response to sound tones in cat auditory cortex neurons: interaction of excitatory and inhibitory effects. *Neuroscience* **43**, 307–321 (1991).
- Ojima, H. & Murakami, K. Intracellular characterization of suppressive responses in supragranular pyramidal neurons of cat primary auditory cortex *in vivo*. *Cereb. Cortex* **12**, 1079–1091 (2002).
- Mehta, M. R., Quirk, M. C. & Wilson, M. A. Experience-dependent asymmetric shape of hippocampal receptive fields. *Neuron* **25**, 707–715 (2000).
- Engert, F., Tao, H. W., Zhang, L. I. & Poo, M. Moving visual stimuli rapidly induce direction sensitivity of developing tectal neurons. *Nature* **419**, 470–475 (2002).
- Rao, R. P. N. & Sejnowski, T. J. In *Advances in Neural Information Processing Systems 12* (eds Solla, S. A., Leen, T. K. & Muller, K. R.) 164–170 (MIT Press, Cambridge, Massachusetts, 2000).
- Zhang, L. I., Bao, S. & Merzenich, M. M. Persistent and specific influences of early acoustic environments on primary auditory cortex. *Nature Neurosci.* **4**, 1123–1130 (2001).
- Zhang, L. I., Tao, H. W., Holt, C. E., Harris, W. A. & Poo, M. A critical window for cooperation and competition among developing retinotectal synapses. *Nature* **395**, 37–44 (1998).
- Moore, C. I. & Nelson, S. B. Spatio-temporal subthreshold receptive fields in the vibrissa representation of rat primary somatosensory cortex. *J. Neurophysiol.* **80**, 2882–2892 (1998).
- Zhu, J. J. & Connors, B. W. Intrinsic firing patterns and whisker-evoked synaptic responses of neurons in the rat barrel cortex. *J. Neurophysiol.* **81**, 1171–1183 (1999).
- Margrie, T. W., Brecht, M. & Sakmann, B. *In vivo*, low-resistance, whole-cell recordings from neurons in the anaesthetized and awake mammalian brain. *Pflügers Arch.* **444**, 491–498 (2002).
- Borg-Graham, L. J., Monier, C. & Fregnac, Y. Visual input evokes transient and strong shunting inhibition in visual cortical neurons. *Nature* **393**, 369–373 (1998).
- Hirsch, J. A., Alonso, J. M., Reid, R. C. & Martinez, L. M. Synaptic integration in striate cortical simple cells. *J. Neurosci.* **18**, 9517–9528 (1998).
- Anderson, J. S., Carandini, M. & Ferster, D. Orientation tuning of input conductance, excitation, and inhibition in cat primary visual cortex. *J. Neurophysiol.* **84**, 909–926 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This work was supported by the Howard Hughes Medical Institute, the John C. and Edward Coleman Fund, the MacDonnell Foundation, the National Institutes of Health and the Sandler Fund.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to L.I.Z. (lizhang@phy.ucsf.edu).

Glutamate-receptor-mediated encoding and retrieval of paired-associate learning

M. Day*, R. Langston & R. G. M. Morris

Laboratory for Cognitive Neuroscience, Division and Centre for Neuroscience, The University of Edinburgh, 1 George Square, Edinburgh EH8 9JZ, UK

Paired-associate learning is often used to examine episodic memory in humans¹. Animal models include the recall of food-cache locations by scrub jays² and sequential memory^{3,4}. Here we report a model in which rats encode, during successive sample trials, two paired associates (flavours of food and their spatial locations) and display better-than-chance recall of one item when cued by the other. In a first study, pairings of a particular foodstuff and its location were never repeated, so ensuring unique 'what-where' attributes. Blocking N-methyl-D-aspartate receptors in the hippocampus—crucial for the induction of certain forms of activity-dependent synaptic plasticity^{5,6}—impaired memory encoding but had no effect on recall. Inactivating hippocampal neural activity by blocking α -amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) receptors impaired both encoding and recall. In a second study, two paired associates were trained repeatedly over 8 weeks in new pairs, but blocking of hippocampal AMPA receptors did not affect their recall. Thus we conclude that unique what-where paired associates depend on encoding and retrieval within a hippocampal memory space^{7,8}, with consolidation of the memory traces representing repeated paired associates in circuits elsewhere.

Paired-associate learning typically consists of a study phase of pairs of items (such as word pairs) followed by a test of cued recall. Our new protocol involves training rats in a large arena containing a 7 × 7 grid of 49 sand-wells that could be uncovered selectively (Fig. 1). A 'sample' trial (Fig. 1a) consists of the rat leaving one of the start boxes to search for a single uncovered sand-well where it can dig to get a 1-g pellet of flavoured rat chow. A second sample trial is given 2 min later with a different flavour of chow at a different location (Fig. 1b). The following cued-recall trial is given in several ways, but on training trials as a rewarded choice. The recall cue is a 500-mg pellet of one of the two earlier flavours, available in a different start box for a 30-second recollection interval during which it is eaten. The animal is then allowed into the arena, which now has two sand-wells uncovered (L1 and L2; Fig. 1c). If cued with food F1, the rat is rewarded with a 1-g pellet of F1 by going to and digging at L1; if cued with F2, it is rewarded with F2 at L2 (see Supplementary Information on counterbalancing). On the next day, a new pair of flavours and spatial locations is chosen as paired associates with an inevitable reuse of both flavours and locations in novel combinations (of which there are 28).

After arena habituation, the rats ($n = 8$) were digging readily for the buried food. The main training then began and, on the measure of first chosen sand-well during a choice trial, better-than-chance performance was observed very early. The rats generally kept their heads near the floor of the arena during sample trials, displaying small lateral head movements reflecting their search for the single, unobtrusive sand-well that was open (see Sample 2 of the Quick-Time movie at <http://neuroweb-2.dns.ed.ac.uk/video/>). On choice trials, the rats ran relatively directly towards one of the two sample sand-wells, indicating recall of location, without displaying these head movements. After only 6 days of training, the average first-

* Present address: GlaxoSmithKline, Centre for Excellence in Drug Discovery, Harlow, Essex, CM19 5AW, UK.

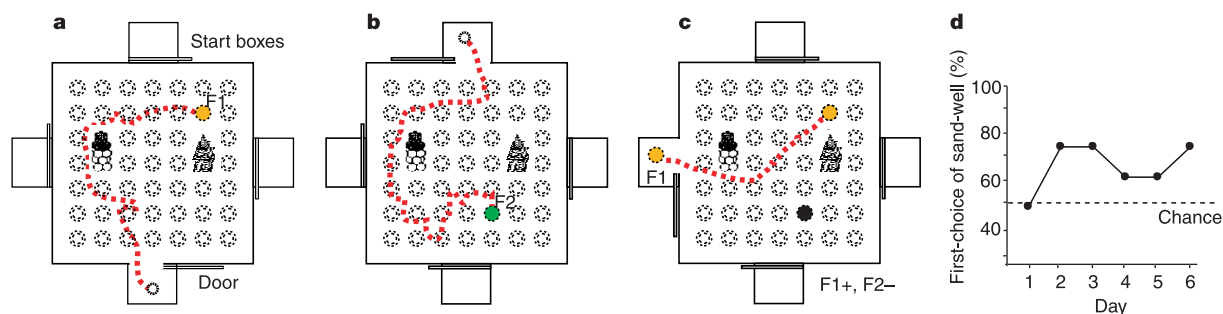


Figure 1 The event arena. **a**, A rat runs into the arena (dotted line), where it displays lateral head movements to find food 1 (F1, orange) at the single open sand-well. **b**, Sample 2 is a different food (F2, green) at a different location. **c**, The cued-recall choice trial begins with presentation of either of the two sample-trial foods (for example, F1 as

shown) with the rat selectively rewarded for digging at the sand-well containing this same food. **d**, Rapid learning over 6 days of training. The rats ($n = 8$) performed better than chance (dotted line) on days 2–6 ($t = 2.65$, d.f. 7, $P < 0.03$).

choice performance over the last 5 days was $70.0 \pm 7.6\%$ (s.e.m.) correct (Fig. 1d), comparable to what is seen in spontaneous alternation.

To check that apparent recall was not mediated by odours emanating from the flavoured foodstuffs, two non-rewarded probe tests were scheduled in which no food was hidden in either of the two sampled sand-wells on choice trials. The animals, together with others first trained with double-sample trials, were trained with the single-presentation sample procedure for a further 10 days of training, and reduced to the most consistent 12 of the original 16 rats ($n = 6$ from each subgroup). The probe tests also offered the opportunity to examine what would happen if a further three sand-wells (locations 3–5) were uncovered that had not been open on the two sample trials of that day. The animals were allowed to dig for up to 60 seconds and we observed above-average digging time at the cued location (Fig. 2a). In a second probe test, the rats were 'misued' by giving them a novel flavour (F3) as a recall cue at the start of the choice trial. The rats now showed equal preference for the two previously sampled locations (L1, L2) that was significantly higher than for the unsampled locations (Fig. 2b). We conclude that rats can encode a flavour–location association in a single trial and later display cued recall of the paired association.

This new one-trial paradigm was explicitly designed to investigate whether activation of glutamate receptors⁹ in the hippocampus would differentially mediate encoding and retrieval memory processes. To do this, we examined the selective antagonists 6-cyano-7-nitroquinoxaline (CNQX, an antagonist of AMPA receptors) and D(–)-2-amino-5-phosphonopentanoic acid (D-AP5, an antagonist of NMDA (*N*-methyl-D-aspartate) receptors) by infusing them (or artificial cerebrospinal fluid; aCSF) into the dorsal hippocampus 15 min before sample trials or 15 min before choice trials (sample and choice were scheduled 20 min apart). A within-subjects protocol was used in which each of the 12 rats was trained over 17 days with 11 regular training days (some of which included infusions with aCSF) interspersed with the six drug-infusion conditions (given as non-rewarded probe trials; see the representative sequence in Fig. 3c).

Rats given aCSF infusions showed better-than-chance choice performance irrespective of when the infusions occurred (Fig. 3c, d). Infusion of D-AP5 impaired choice accuracy when infused before sample trials, but not when infused after sample trials but before choice trials. Infusion of CNQX impaired choice accuracy whether given before or after a sample trial. This dissociation was by no means a foregone conclusion, but makes sense considering the role of NMDA receptors in the induction of activity-dependent synaptic plasticity in the hippocampus¹⁰ (Fig. 3b) and of AMPA-mediated fast synaptic transmission in expressing changes in

synaptic weight¹¹ (Fig. 3a, ref. 11).

There are two caveats, both related to the involvement of the hippocampus in spatial memory^{12,13}. One is that NMDA receptors might be necessary for encoding the place of a novel sand-well within a familiar environment, as in a water maze¹⁴, and this encoding might be a prerequisite for forming a flavour–place association. The other is that the deficit in paired-associate recall during hippocampal inactivation by CNQX might be secondary to a disruption of spatial memory. The latter caveat is unlikely because the animals had received extensive exposure to the arena by the time

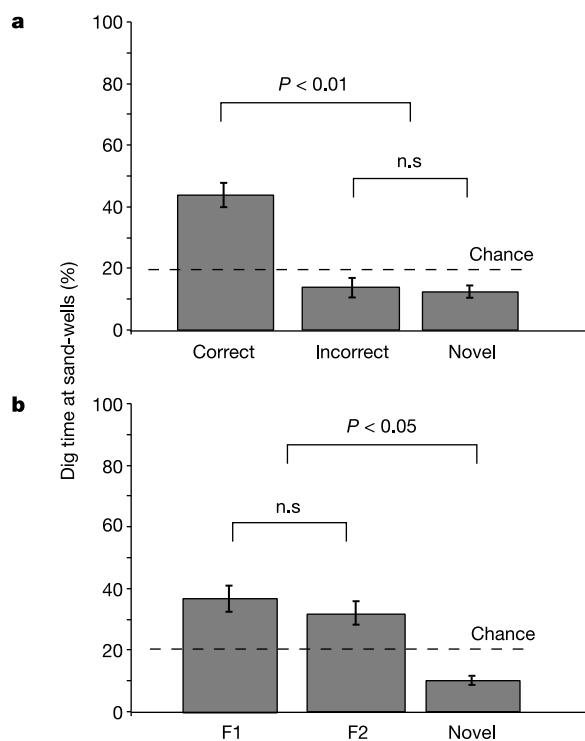


Figure 2 Non-rewarded probe tests. **a**, Percentage of time spent digging on choice trials was significantly different at the correct (cued), incorrect and novel sand-wells ($F = 6.92$, d.f. 2/22, $P < 0.005$). Orthogonal comparisons revealed that the time spent at the incorrect and novel locations did not differ ($F < 1$), but time spent at the correct location was significantly higher ($F = 13.8$, d.f. 1/22, $P < 0.01$). **b**, Choice trials 'misued' by a flavour not used during the sample trials revealed that time spent at the two sampled locations did not differ ($F < 1$) but was significantly higher than at the novel locations ($F = 5.27$, d.f. 1/22, $P < 0.05$).

of the drug infusions (2 months), sufficient to allow spatial information to be consolidated outside the hippocampus¹⁵. We could exclude this second concern if memory retrieval can then bypass this structure. Two flavour–place pairings (apple at location (1,5); brandy at location (7,5)) were therefore trained repeatedly over 8 weeks, amid further training on single-trial novel associates. Both paired associates had at least 20 training presentations before the start of non-rewarded probe tests. The training protocol included two sample trials and one choice trial each day, as before, but choice trials in which a total of four sand-wells were generally available (see Supplementary Information). Non-rewarded probe trials included the cued sand-well; a non-cued but previously sampled sand-well; a never rewarded sand-well (always at (3,3)); and a novel well (Fig. 4a).

The two repeat pairings were gradually learned over the 2-month training period (for example, an increasing proportion of time spent digging at the relevant wells on rewarded choice trials: $F = 3.67$, d.f. 2/24, $P < 0.05$; data not shown). The key tests of whether cued recall of ‘consolidated’ associates would be possible during hippocampal inactivation were conducted in a counter-balanced manner over the next 2 weeks (weeks 9 and 10), interspersed with further training days to maintain performance, with a further probe trial in week 11. Intrahippocampal CNQX had no effect on the ability of rats to navigate appropriately and dig persistently at the correct location when cued with either repeat-pair flavour (Fig. 4b). The final probe trials (data not shown) confirmed that the animals could be cued by the relevant flavour to choose, under CNQX or CSF, between just the two repeat-trial

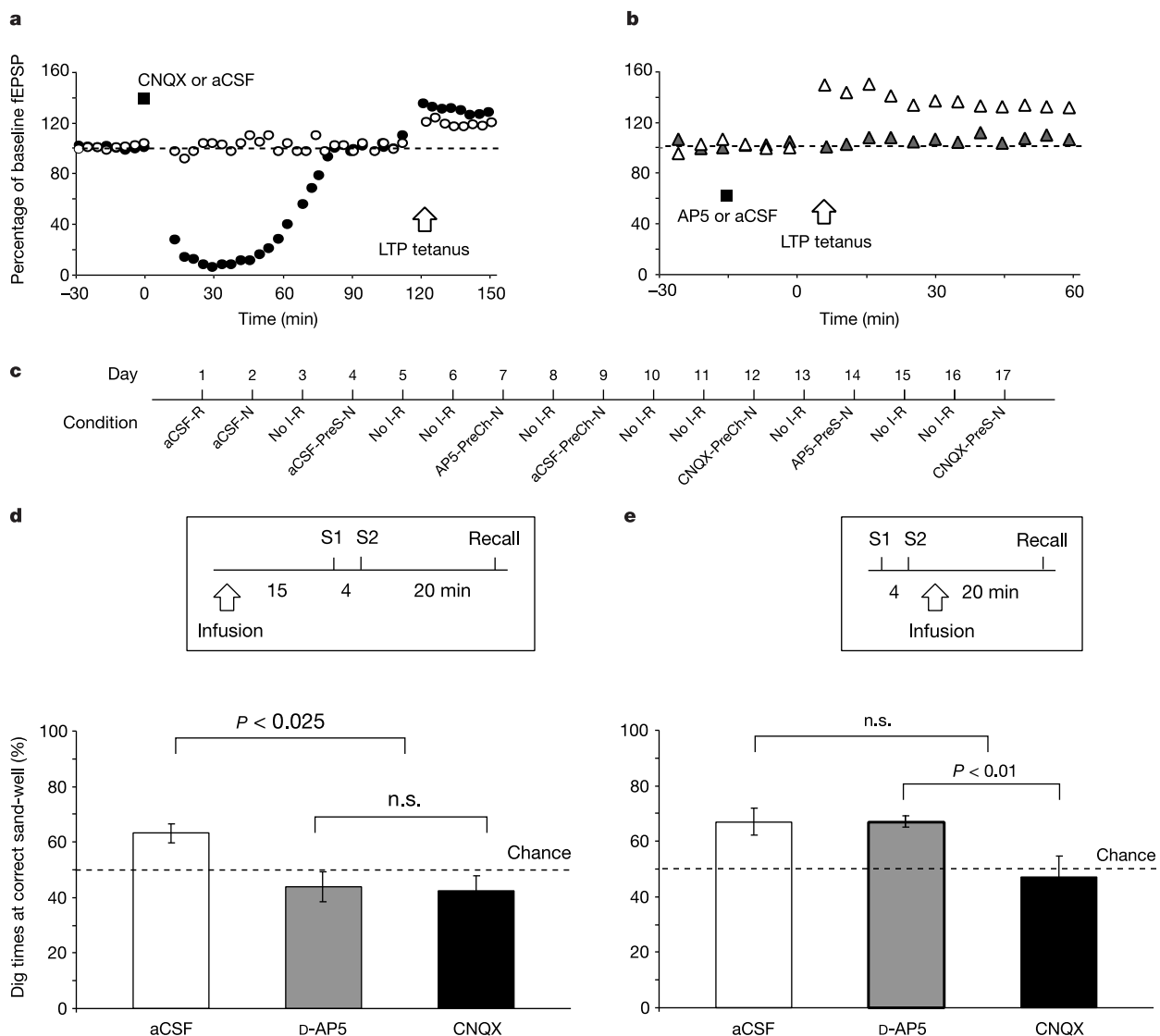


Figure 3 Differential glutamate-receptor dependence of encoding and retrieval. **a**, Maximal neural inactivation in the hippocampus occurred within 10–15 min of CNQX infusion and lasted for ~60 min. Open circles, aCSF; filled circles, CNQX. LTP, long-term potentiation. **b**, D-AP5 infusions did not affect fast synaptic transmission but blocked LTP induction 15 min after infusion. Open triangles, aCSF; filled triangles, D-AP5. **c**, Representative sequence of drug treatments for an individual rat across 17 days (counterbalanced with respect to treatment order). D-AP5, CNQX and aCSF indicate bilateral infusion; R and N indicate that a choice trial was rewarded or non-rewarded; PreS and PreCh indicate infusions before sample or choice. Nol signifies no infusion.

d, Drug-infusions before sample trials revealed an overall effect of drug ($F = 4.61$, d.f. 2/20, $P < 0.025$). There was better-than-chance and above-chance performance on aCSF days than on drug days ($F = 5.30$, d.f. 1/20, $P < 0.01$; $t = 3.86$, d.f. 10, $P < 0.003$), and no difference between D-AP5 and CNQX treatment days ($F < 1$), which were both at chance. **e**, Drug infusions before choice trials also revealed an overall drug effect ($F = 4.38$, d.f. 2/20, $P < 0.03$), no difference between days with aCSF and D-AP5 treatment ($F < 1$), but better-than-chance and above-chance performance on these days compared with those with CNQX infusions ($F = 10.76$, d.f. 1/20, $P < 0.01$; aCSF $t = 3.87$, d.f. 10, $P < 0.003$; D-AP5 $t = 4.7$, d.f. 10, $P < 0.001$).

locations with performance significantly above chance ($t = 3.73$, d.f. 1/11, $P < 0.003$). We also routinely conducted the 'positive control' of establishing electrophysiological inactivation of the hippocampus in other rats by using the same vial of CNQX as used behaviourally. It follows that the animals are not 'lost in space' when the hippocampus is inactivated and can retrieve the spatial locations of repeat-trial cued flavours.

Our experiments establish that rats can, in one trial, encode a novel association between a food flavour ('what') and a spatial location ('where') and can preferentially recall the correct location when cued with the flavour item of the paired associate. Previous studies of paired-associate learning in animals have used multi-trial training procedures and recognition tests^{13,16,17}. Here, cued recall after only one encoding trial is not an artefact of cryptic odour guidance, because performance was above chance during non-rewarded probe tests. Choice cannot be guided by recognition memory or relative recency as in the widely used task involving delayed non-matching to sample^{18,19}, because the two flavours and locations used as samples on a given day were arranged to be equally familiar and counterbalanced with respect to order. Any differential preference for one flavour over the other was fully controlled by counterbalancing, with the miscue test establishing that the rats would choose equally between the two samples when cued with a different flavour. Thus, the event arena offers a valid test of

associative cued recall in animals ('what-where' in the terminology of ref. 2).

Our neurobiological findings indicate that blocking hippocampal NMDA receptors during sample trials impairs the memory encoding of paired associates but has no effect on retrieval. We suspect that the effect on encoding is primary, with information about the 'what' and 'where' of single events associated in hippocampal circuitry, although a cascade of NMDA-receptor-dependent components of event memory might be involved. Earlier studies with a strictly spatial paradigm emphasized the event-like nature of one-trial spatial memory¹⁴. Pairing of events with familiar places could work if recall of the spatial information, from a neocortical site of storage, was by means of the direct input to CA1 from the entorhinal cortex²⁰, with event processing upstream in hippocampal circuitry. Blocking hippocampal fast synaptic transmission impairs both encoding and retrieval, a finding compatible with activations seen during human functional imaging²¹ and experience-induced strengthening of transmission by means of AMPA-receptor insertion²². However, the interval between sample and recall trials was too short to realize a firm dissociation. Finally, if a flavour-place pair is trained repeatedly, accurate retrieval is still possible when the hippocampus is inactivated. The rats will most probably do this by accessing a consolidated neocortical memory trace of the paired associate rather than by recalling the events of the earlier sample trials by means of the now-inaccessible hippocampal trace. Speculatively, this is akin to the distinction in human studies between 'knowing' and 'remembering'²³, but a literal application of this distinction to animals is premature. The protocol's flexibility could make it valuable as an animal analogue of a diagnostic test for mild Alzheimer's disease²⁴ and to analyse the neural mechanisms of episodic-like memory including, in due course, the incorporation of temporal and sequential factors^{3,4,25}. □

Methods

Subjects

Male Lister hooded rats (initially 3 months old) were maintained on water *ad libitum* but a restricted diet to maintain a body weight at 85% of the free-feeding weight. Procedures complied with the UK Animals (Scientific Procedures) Act, PPL 60/2484.

Apparatus

The arena was made of clear Perspex (1.6 m × 1.6 m, 0.3 m side walls). The four start boxes (0.3 m × 0.3 m) were placed centrally in each wall, with sliding doors for arena access. The 49 food wells (7 cm diameter, 4 cm deep) were normally covered by an insert made of white Perspex, the same colour as the floor, to ensure an unbroken surface. Two unused locations (row 4, column 2—(4,2)—and (4,6)) were covered by distinctive landmarks 20 cm high (a pagoda and a glued stack of golf balls). The sand inside each sand-well was slightly adulterated by ground-up rat chow of all flavours, thoroughly stirred, and renewed regularly. One or more sand-wells could be exposed on sample or choice trials and, if rewarded, contained a single 1-g pellet of flavoured rat chow. The 28 flavours available included: a base flavour; single flavours such as almond, anise and apple; and combination flavours (manufactured as a single pellet) such as banana plus cherry (full list in Supplementary Information).

Training procedure

The animals were observed by means of two video cameras connected to video recorders and computer software for tracking (videos of the animals performing the task are available at <http://neuroweb-2.dns.ed.ac.uk/video/>).

In habituation experiments, on each of five days the rats explored the arena, initially found food on top of open sand-wells and then learned to dig for food pellets that were gradually hidden under the sand.

Training was as described in the text for 5 days per week; two flavours were used on sample trials and, typically, one only of these two flavours was available at the choice trial.

In experiment 1, different flavours were used on successive days and, when reused after 10–15 days, were always in different flavour-location combinations. Sample trials were rewarded with 1-g food pellets. The reminder cue in the start box on choice trials was 500 g and choice trials were rewarded with 1-g food pellets (allowing correction). Training continued over 8 months. The first phase of training monitored first choices on rewarded cued-recall choice trials ($n = 8$) preceded by single-presentation or double-presentation sample trials (see Supplementary Information). We then switched all rats ($n = 16$, using only the 12 rats that had learned the task) to single-presentation sample trials for 15 days before interspersing non-rewarded probe tests with rewarded choice trials. These probe tests permitted measures of digging times on both cued-recall and miscued-recall choice trials with five dig locations. Digging times were averaged over two non-rewarded days, enabling counterbalancing with respect to the sample flavour that was cued at recall. In

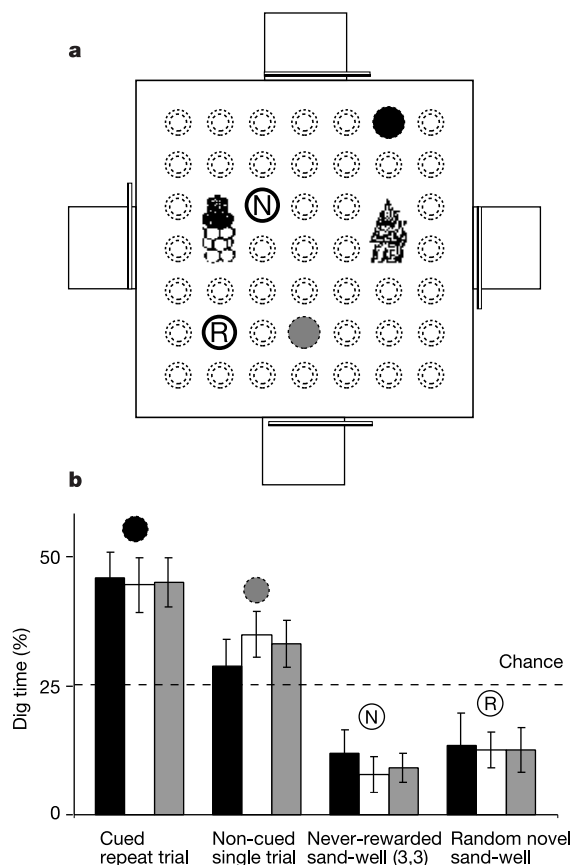


Figure 4 Insensitivity of repeat-trial paired associates to hippocampal inactivation. **a**, Event arena with representative layout of four sand-wells used during non-rewarded probe tests: a cued repeat-trial location (for example (1,5); black circle), a non-cued single-trial location (grey circle), the never-rewarded location, (3,3) (N), and a random novel location (R). **b**, Different proportions of digging were observed at the four sand-wells ($F = 26.1$, d.f. 2/24, $P < 0.0001$), but intrahippocampal CNQX (black bars) did not disrupt performance ($F < 1$). The cued repeat-trial location was above chance ($P < 0.005$); the never-rewarded location was below chance ($P < 0.005$). White bars, aCSF; grey bars, no infusion.

examining blocking of hippocampal AMPA and NMDA receptors with CNQX, D-AP5 and aCSF, drugs were applied 15 min before sample trials or 15 min before choice trials (one day for each of six conditions), with a 20-min interval between sample and choice trials. In-house software (LabView) was used to monitor digging.

Experiment 2 ($n = 12$) was conducted with new animals, with two flavour–place pairings repeated amid other novel pairs. After habituation, the rats were given extended training (5 days per week for 8 weeks): In any week, one day involved sample presentations and then a choice trial of the repeated flavour–location pairs (apple in location (1,5); brandy in location (7,5)); one day only novel pairs; and three days with mixed pairs (for example a novel pair for sample 1 and apple in (1,5) for sample 2, or brandy, or vice versa). Over 8 weeks, the animals received 20 sample presentations of the repeat-trial pairs (8×2.5). The choice trials always had four sand-wells open: the cued location, the non-cued but previously sampled location, a never-rewarded location, (3,3), and a novel location. Probe tests followed during weeks 9–12 of testing, including CNQX, aCSF or no-drug conditions, these being interspersed with further training days. Four sand-wells were also used in these probe tests (see Fig. 4).

Surgery, electrophysiology and drug infusions

Stainless-steel guide cannulae (26-gauge) were implanted into the dorsal hippocampus bilaterally using standard stereotaxic techniques under tribromoethanol anaesthesia. Electrophysiological studies to calculate the CNQX and D-AP5 doses used separate animals under urethane anaesthesia (1.5 g kg^{-1}). Dentate field potentials were monitored unilaterally by means of a recording electrode ($75 \mu\text{M}$) lowered into the hilus, with a bipolar stimulating electrode placed into the angular bundle of the perforant path. The concentric bipolar recording electrode (Rhodes) was $1,400 \mu\text{m}$ from the infusion cannulae. The slope, amplitude and population spike of the field potentials was monitored with in-house software (LabView). Drug infusions in these and the behavioural experiments were $1 \mu\text{l}$ CNQX (3 mM ; Tocris) and $1 \mu\text{l}$ D-AP5 (30 mM ; Tocris) per cannula. The infusion cannula was fixed for the electrophysiological experiments, but was left in place for 1 min after infusion for the behavioural experiments, with the animal lightly restrained in the hand of one of the experimenters (M.D. or R.L.).

Received 17 December 2002; accepted 29 April 2003; doi:10.1038/nature01769.

1. Tulving, E. *Elements of Episodic Memory* (Oxford Univ. Press, New York, 1983).
2. Clayton, N. S. & Dickinson, A. Episodic-like memory during cache recovery by scrub jays. *Nature* **395**, 272–274 (1998).
3. Fortin, N. J., Agster, K. L. & Eichenbaum, H. B. Critical role of the hippocampus in memory for sequences of events. *Nature Neurosci.* **5**, 458–462 (2002).
4. Kesner, R. P., Gilbert, P. E. & Barua, L. A. The role of the hippocampus in memory for the temporal order of a sequence of odors. *Behav. Neurosci.* **116**, 286–290 (2002).
5. Bliss, T. V. P. & Lomo, T. Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *J. Physiol. (Lond.)* **232**, 331–356 (1973).
6. Martin, S. J., Grimwood, P. D. & Morris, R. G. M. Synaptic plasticity and memory: An evaluation of the hypothesis. *Annu. Rev. Neurosci.* **23**, 649–711 (2000).
7. Morris, R. G. M. & Frey, U. Hippocampal synaptic plasticity: Role in spatial learning or the automatic recording of attended experience? *Phil. Trans. R. Soc. Lond. B* **352**, 1489–1503 (1997).
8. Eichenbaum, H., Dudchenko, P. A., Wood, E. R., Shapiro, M. & Tanila, H. The hippocampus, memory and place cells: Is it a spatial memory or a memory space? *Neuron* **23**, 209–226 (1999).
9. Bleakman, D. & Lodge, D. Neuropharmacology of AMPA and kainate receptors. *Neuropharmacology* **37**, 1187–1204 (1998).
10. Bliss, T. V. & Collingridge, G. L. A synaptic model of memory: Long-term potentiation in the hippocampus. *Nature* **361**, 31–39 (1993).
11. Cowan, W. M., Sudhof, T. C. & Stevens, C. F. *Synapses* (Johns Hopkins Univ. Press, 2001).
12. O'Keefe, J. & Nadel, L. *The Hippocampus as a Cognitive Map* (Clarendon Press, Oxford, 1978).
13. Gilbert, P. E. & Kesner, R. P. Role of the rodent hippocampus in paired-associate learning involving associations between a stimulus and a spatial location. *Behav. Neurosci.* **116**, 63–71 (2002).
14. Steele, R. J. & Morris, R. G. M. Delay-dependent impairment of a matching-to-place task with chronic and intrahippocampal infusion of the NMDA-antagonist D-AP5. *Hippocampus* **9**, 118–136 (1999).
15. Kim, J. J. & Fanselow, M. S. Modality-specific retrograde amnesia of fear. *Science* **256**, 675–677 (1992).
16. Sakai, K. & Miyashita, Y. Neural organization for the long-term memory of paired associates. *Nature* **354**, 152–155 (1991).
17. Gaffan, D. Scene-specific memory for objects: A model of episodic memory impairment in monkeys with fornix transection. *J. Cogn. Neurosci.* **6**, 305–320 (1994).
18. Mishkin, M. & Delacour, J. An analysis of short-term visual memory in the monkey. *J. Exp. Psychol.: Anim. Behav. Processes* **1**, 326–334 (1975).
19. Zola-Morgan, S. & Squire, L. R. Medial temporal lesions in monkeys impair memory on a variety of tasks sensitive to human amnesia. *Behav. Neurosci.* **99**, 22–34 (1985).
20. Brun, V. H. et al. Place cells and place recognition maintained by direct entorhinal–hippocampal circuitry. *Science* **296**, 2243–2246 (2002).
21. Zeineh, M. M., Engel, S. A., Thompson, P. M. & Bookheimer, S. Y. Dynamics of the Hippocampus during encoding and retrieval of face-name pairs. *Science* **299**, 577–580 (2003).
22. Takahashi, T., Svoboda, K. & Malinow, R. Experience strengthening transmission by driving AMPA receptors into synapses. *Science* **299**, 1585–1588 (2003).
23. Gardiner, J. M. in *Episodic Memory: New Directions in Research* (eds Baddeley, A., Conway, M. & Aggleton, J.) 11–30 (Oxford Univ. Press, 2001).
24. Swainson, R. et al. Early detection and differential diagnosis of Alzheimer's disease and depression with neuropsychological tasks. *Dement. Geriatr. Cogn. Disord.* **12**, 265–280 (2001).
25. Lisman, J. E. Relating hippocampal circuitry to function: Recall of memory sequences by reciprocal dentate–CA3 interactions. *Neuron* **22**, 233–242 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank N. Clayton, D. Donaldson, P. Dudchenko, L. Jacobs, E. Moser and E. Wood for discussion. Funded by MRC Innovation and Programme Grants and an E.U. Framework V Grant.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.G.M.M. (r.g.m.morris@ed.ac.uk).

T-type calcium channel regulation by specific G-protein $\beta\gamma$ subunits

Joshua T. Wolfe, Hongge Wang, Jason Howard, James C. Garrison & Paula Q. Barrett

Department of Pharmacology, University of Virginia, Charlottesville, Virginia 22908, USA

Low-voltage-activated (LVA) T-type calcium channels have a wide tissue distribution and have well-documented roles in the control of action potential burst generation and hormone secretion¹. In neurons of the central nervous system and secretory cells of the adrenal and pituitary, LVA channels are inhibited by activation of G-protein-coupled receptors that generate membrane-delimited signals^{2–5}, yet these signals have not been identified. Here we show that the inhibition of α_{1H} ($\text{Ca}_v3.2$), but not α_{1G} ($\text{Ca}_v3.1$) LVA Ca^{2+} channels is mediated selectively by $\beta_2\gamma_2$ subunits that bind to the intracellular loop connecting channel transmembrane domains II and III. This region of the α_{1H} channel is crucial for inhibition, because its replacement abrogates inhibition and its transfer to non-modulated α_{1G} channels confers $\beta_2\gamma_2$ -dependent inhibition. $\beta\gamma$ reduces channel activity independent of voltage, a mechanism distinct from the established $\beta\gamma$ -dependent inhibition of non-L-type high-voltage-activated channels of the Ca_v2 family^{6,7}. These studies identify the α_{1H} channel as a new effector for G-protein $\beta\gamma$ subunits, and highlight the selective signalling roles available for particular $\beta\gamma$ combinations.

G-protein $\beta\gamma$ subunits released by receptor activation have important physiological functions in the regulation of G-protein-gated, inward-rectifier potassium (GIRK)⁸; and $\text{Ca}_v2 \text{ Ca}^{2+}$ (refs 8, 9) channels (N-type, P/Q-type and R-type), by interacting with intracellular domains of the channel molecules. In the Ca_v2 family, these domains include the I–II cytoplasmic loop and the amino and carboxy termini^{10–12}. The effects of $\beta\gamma$ inhibition on the Ca_v2 family members are also modulated by protein kinase C (ref. 13) and the synaptic proteins syntaxin 1A, SNAP-25 and cysteine-string protein¹⁴. By contrast, neither the identity of the G-protein subunit mediating the inhibition of LVA channels nor the domain of the LVA channel supporting inhibition is known.

We tested the possibility that G-protein $\beta\gamma$ subunits could regulate whole cell LVA currents, comparing the effects of $\beta_2\gamma_2$ and $\beta_1\gamma_2$ overexpression in HEK-293 cells stably expressing $\alpha_{1H} \text{ Ca}^{2+}$ channels (Fig. 1a, upper panels). At all potentials tested, α_{1H} channels are specifically inhibited by $\beta_2\gamma_2$ dimers, decreasing current density by 50% (Fig. 1b, left panel). Surprisingly, the $\beta_1\gamma_2$ dimer did not inhibit the channel. This difference in inhibitory selectivity between the two β subunits could not be attributed to inadequate levels of $\beta_1\gamma_2$ protein expression. Immunoblots of the transfected cell lysates (Fig. 1b, inset) show that the expression of green fluorescent protein (GFP), a marker for transfection efficiency, is similar in both cases, and that each Flag-tagged β subunit

is expressed at comparable levels. Although the expression of $\beta_3\gamma_2$ or $\beta_4\gamma_2$ was less robust, the α_{1H} currents in cells expressing these two $\beta\gamma$ isoforms were not inhibited and were similar to those seen with GFP expression alone (Fig. 1c), indicating that the inhibition of α_{1H} channels is dependent on the subtype of $\beta\gamma$ dimer. Specificity in $\beta\gamma$ function has also been observed with N-type Ca^{2+} channels, because all $\beta\gamma$ dimers can inhibit channel activity¹⁵, but only inhibition mediated by β_1 -containing dimers can be antagonized by protein kinase C (ref. 16).

Because members of the LVA Ca^{2+} channel family share only 40% homology, we investigated whether $\beta\gamma$ dimers could also inhibit Ca^{2+} channel currents in HEK cells stably expressing another member of the Ca_v3 family, the α_{1G} channel (Fig. 1a, lower panels). Under identical recording conditions, $\beta_1\gamma_2$ or $\beta_2\gamma_2$ was unable to decrease whole-cell α_{1G} currents (Fig. 1b, right panel), despite expression at levels that support the inhibition of α_{1H} channels. These findings of channel-subtype-specific regulation by $\beta\gamma$ are unique to the Ca_v3 family, because all members of the Ca_v2 family

can be regulated by $\beta\gamma$ dimers¹⁵. To determine the relative importance of the γ subunit to the observed specificity of $\beta\gamma$ action, we first tested β_2 subunits dimerized with one member from each of the five γ -subunit families¹⁷. Four β_2 -containing dimers (with γ_2 , γ_{10} , γ_{12} and γ_{13}) inhibited current (Fig. 1c), despite high structural divergence between the γ subunits (up to 80%). Only when paired with γ_{11} , which contains the farnesyl isoprenoid group¹⁷, did β_2 fail to inhibit α_{1H} current. Subsequently, we chose γ_{10} , which shares only 50% similarity with γ_2 , to dimerize with each of the other β subunits (β_1 , β_3 , β_4 ; Fig. 1c). None of these three β subunits was effective when dimerized with either γ_2 or γ_{10} . In addition, we corroborate previously published findings¹⁸ and show that $\beta_1\gamma_2$ (Fig. 1d), $\beta_3\gamma_2$ and $\beta_4\gamma_2$ (not shown) stimulate current from GIRK1,4 channels as effectively as $\beta_2\gamma_2$. These data show that the $\beta\gamma$ selectivity we have observed is effector specific and not the result of formation of non-functional complexes. Collectively, these data indicate that the selectivity of current inhibition is conferred by the β subunit, and is specific to β_2 .

$\beta\gamma$ -subunit binding to Ca_v2 family members induces a 20-mV depolarizing shift in the half-activation potential ($V_{1/2}$ of activation)⁶. To assess the voltage dependence of inhibition of α_{1H} channels by $\beta_2\gamma_2$, we determined the voltage dependences of activation and steady-state inactivation using isochronal tail-current protocols. Expression of $\beta_2\gamma_2$ proportionally decreased current at all voltages, and thus did not change the dependence of activation or inactivation on voltage (Fig. 2a, b), indicating that modulation of LVA channels by $\beta\gamma$ proteins is independent of voltage. Inhibition of Ca_v2 channels by $\beta\gamma$ proteins is also temporarily relieved by strong depolarizing voltage pulses of short duration (facilitation) that presumably release $\beta\gamma$ proteins from the channel structure^{7,9}. LVA Ca^{2+} channels inactivate more rapidly, to a greater extent, and recover more slowly from inactivation than high-voltage-activated (HVA) channels, characteristics that could combine to obscure prepulse facilitation¹⁹. We compared peak current amplitudes and recovery from inactivation at various times (1–2,000 ms) after a 100-ms conditioning pulse to +80 mV. In the absence of $\beta\gamma$ expression, application of a strong depolarizing prepulse decreased α_{1H} channel currents; in $\beta_2\gamma_2$ -expressing cells, prepulse delivery did not facilitate α_{1H} currents (Fig. 2c) or enhance the time course of recovery from inactivation (not shown). Taken together, these results show that modulation of LVA channels by $\beta\gamma$ is independent of voltage and therefore distinct from that of HVA channels.

To determine whether the voltage-independent decrease in channel current mediated by $\beta_2\gamma_2$ results from a decrease in the number of channels at the cell surface²⁰, transfected cells were first selected by fluorescence-assisted cell sorting. GFP-positive cells were compared for cell-surface α_{1H} expression after cell-surface biotinylation followed by precipitation with streptavidin. Immunoblots for α_{1H} in three independent experiments show that $\beta_2\gamma_2$ does not change the amount of channel protein at the plasma membrane (Fig. 2d), indicating that $\beta_2\gamma_2$ decreases the functionality of channels in the membrane.

Direct binding of $\beta\gamma$ to a QXXER motif seems to underlie the regulation by $\beta\gamma$ of some effectors, such as adenylyl cyclase II, GIRK1 and β -adrenergic kinase 1 (ref. 21). Because the amino-acid sequence of α_{1H} channels contains a QXXER motif in the II–III loop (residues 1159–1163), we compared the ability of purified, recombinant $\beta_1\gamma_2$ and $\beta_2\gamma_2$ to bind the α_{1H} II–III loop, expressed and purified as a 47-kDa fusion protein with glutathione S-transferase (GST). The results show that $\beta_2\gamma_2$ binds in a dose-dependent manner to the α_{1H} II–III loop (Fig. 3a, left panel); by contrast, neither $\beta_1\gamma_2$ (right panel), $\beta_3\gamma_2$, nor $\beta_4\gamma_2$ (not shown) bound to the II–III loop. To confirm the finding that $\beta\gamma$ binding is correlated with the inhibition of channel current, we expressed the analogous II–III loop of α_{1G} as a GST fusion protein and assayed the binding of $\beta\gamma$. Consistent with the observation that neither $\beta\gamma$ dimer inhibits α_{1G}

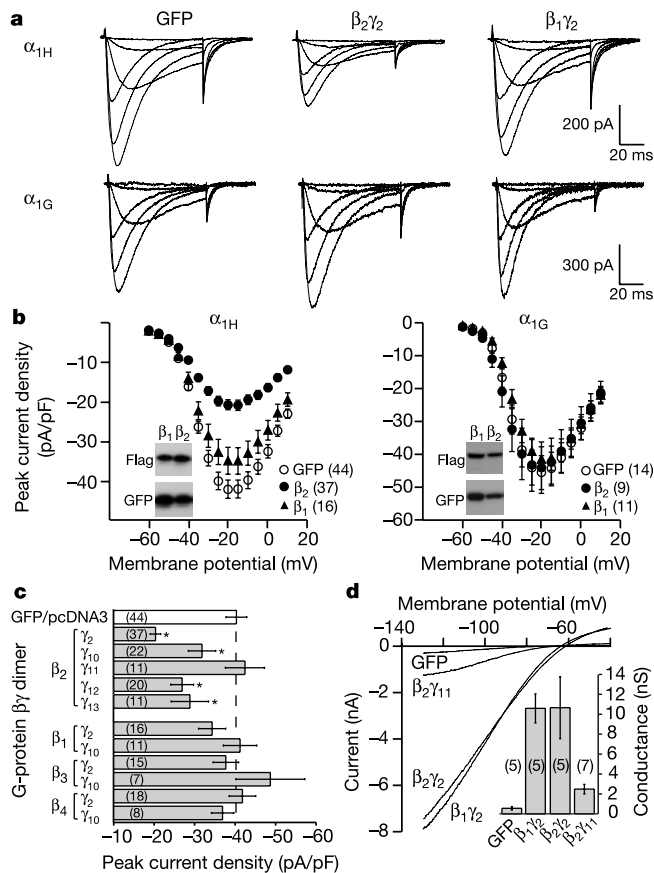


Figure 1 Effect of G-protein $\beta\gamma$ expression on α_{1H} and α_{1G} Ca^{2+} currents. **a**, Sample currents from α_{1H} (top) or α_{1G} (bottom) channels recorded at $V_m = -60, -50, -40, -20, -5$ and $+10$ mV (-90 mV holding potential) in HEK-293 cells expressing marker GFP and either empty vector (left), Flag- $\beta_2\gamma_2$ (middle) or Flag- γ_2 (right). **b**, Effect of $\beta\gamma$ expression on current-voltage relationships for α_{1H} or α_{1G} channels. Current amplitude is expressed as picoamps/picofarads (pA/pF) to correct for differences in cell size. Insets are immunoblots of cell lysates showing similar levels of β -subunit expression; antibodies detected Flag-tagged β subunits (top) or GFP (bottom). The number of cells in each data set is indicated next to each symbol. **c**, Histogram showing the effects of $\beta\gamma$ dimer combinations on peak current density relative to GFP-transfected cells. Asterisk, $P < 0.05$. **d**, GIRK1,4 currents stimulated by transfection of $\beta\gamma$ cDNAs, relative to cells transfected with GFP. The inset shows relative K^+ conductances for the four treatments. The number of cells used in each treatment is indicated in parentheses next to each symbol. Error bars are s.e.m.

channel current, neither $\beta_2\gamma_2$ (Fig. 3b, left) nor $\beta_1\gamma_2$ (Fig. 3b, right) bound to the α_{1G} II–III-loop GST fusion protein.

To test the hypothesis that α_{1H} current inhibition is dependent on $\beta\gamma$ binding to the II–III loop, we generated two chimaeric channels in which the II–III loops of α_{1H} and α_{1G} channels were swapped. Each chimaera was stably expressed in HEK-293 cells and tested for inhibition by $\beta_2\gamma_2$. Replacement of the α_{1H} II–III loop with that from α_{1G} generated a mutant α_{1H} channel ($\alpha_{1H}(G_{II-III})$), with gating properties similar to wild-type α_{1H} channels, that was indeed refractory to inhibition by $\beta_2\gamma_2$. By contrast, α_{1G} channels were rendered sensitive to $\beta_2\gamma_2$ inhibition when the II–III loop was replaced with that from α_{1H} channels (Fig. 3c, d). These functional results, combined with our binding data, provide strong evidence that $\beta_2\gamma_2$ inhibits α_{1H} channels by a mechanism that depends on direct binding to the II–III loop. We also investigated the functional importance of the QXXER motif in $\beta\gamma$ binding by mutating critical residues to alanines (QCGER $\rightarrow$ ACGAA), a strategy that interferes with the binding of $\beta\gamma$ to cyclase²¹. The mutant α_{1H} II–III loop retained the ability to bind $\beta\gamma$ (data not shown), thus excluding the QXXER motif as the direct site of $\beta\gamma$ binding. These findings are

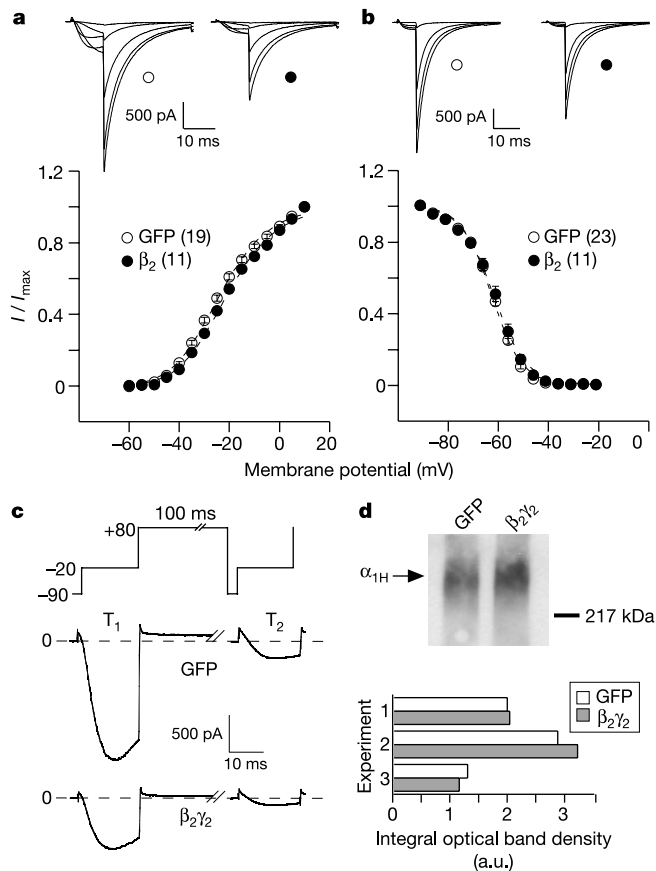


Figure 2 Characterization of channel inhibition induced by $\beta_2\gamma_2$. **a**, Sample currents and voltage dependence of activation plotted as relative conductance (normalized tail-current amplitudes, I/I_{max}) against V_m for α_{1H} channels expressed with (solid circles) or without (open circles) $\beta_2\gamma_2$. **b**, Sample currents and voltage dependence of steady-state inactivation plotted as relative availability (normalized tail-current amplitudes, I/I_{max}) after 6 s against V_{hold} . **c**, Sample traces from a double-pulse protocol as diagrammed, recorded from cells transfected with ($n = 7$) or without ($n = 6$) $\beta_2\gamma_2$. Peak amplitude of T_2 (second test pulse to -20 mV for 18.24 ms) is $16.2 \pm 0.6\%$ of T_1 (first test pulse to -20 mV for 18.24 ms) for control and $17.8 \pm 0.7\%$ of T_1 for $\beta_2\gamma_2$. **d**, Typical immunoblot of α_{1H} channels after cell-surface biotinylation of GFP-positive cells transfected with empty vector or $\beta_2\gamma_2$. The bar graph shows integrated optical band densities in three independent experiments.

consistent with previous observations showing that the QXXER motif does not support the regulation by $\beta\gamma$ of certain effectors, notably HVA Ca^{2+} channels^{11,12,22}.

Because dopamine inhibits LVA Ca^{2+} channels expressed in a variety of cells including the adrenal zona glomerulosa, where the α_{1H} subtype predominates²³, we sought to investigate the importance of $\beta\gamma$ activation and the α_{1H} II–III loop in a receptor-coupled system. We used a human adrenocarcinoma cell line (H295R) that advantageously expresses β_2 at much higher levels than other G-protein β subunits (Fig. 4c). Untransfected H295R cells express minimal endogenous LVA current and are capable of steroid synthesis²⁴. Because primer-specific amplification from total H295R RNA revealed message for D_1 -dopamine but not D_2 -dopamine receptors, we recorded whole-cell currents from α_{1H} -transfected cells and measured the response to SKF 38393, a specific D_1 -type receptor agonist. Transfected channel currents were

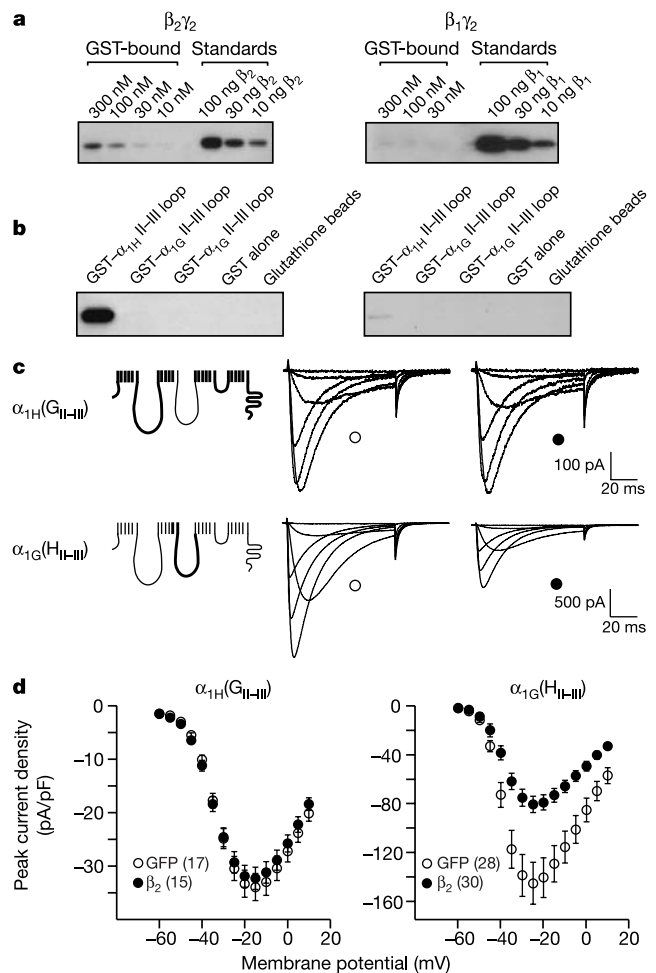


Figure 3 The α_{1H} channel II–III loop supports $\beta_2\gamma_2$ binding and contributes to G-protein-dependent inhibition. **a**, *In vitro* binding of increasing concentrations of purified, recombinant $\beta_2\gamma_2$ (left) or $\beta_1\gamma_2$ (right) at the indicated concentrations, to 200 nM α_{1H} II–III loop-GST fusion protein. Immunoblots with β -specific antibodies detect bound $\beta\gamma$ dimers and recombinant protein standards that serve to quantify binding. **b**, Binding specificity of $\beta_2\gamma_2$ (left) or $\beta_1\gamma_2$ (right) to 200 nM purified GST fusion proteins (α_{1H} II–III loop GST, α_{1G} II–III loop GST, GST alone or glutathione beads) as in **a**. **c**, Channel chimaera diagrams (dark, α_{1H} ; light, α_{1G}) (left) and sample currents from chimaeric channels in the absence (open symbols) or presence (filled symbols) of $\beta_2\gamma_2$ expression (right). **d**, Current–voltage relationships for $\alpha_{1H}(G_{II-III})$ (left) or $\alpha_{1G}(H_{II-III})$ (right) chimaeras, transfected with (filled symbols) or without (open symbols) $\beta_2\gamma_2$. The number of cells in each data set is indicated in parentheses next to each symbol. Error bars are s.e.m.

about 10-fold greater than endogenous currents (Fig. 4a): the observed current was therefore carried almost exclusively by the transfected channel. Activation of the D₁ receptor inhibited α_{1H} currents (25–30%) over the course of 2–3 min of slow drug perfusion (Fig. 4a, b, d) and this effect was antagonized by SCH 23390, a specific blocker of the D₁ receptor (not shown). Inhibition was selective for the α_{1H} channel subtype, because α_{1G} channel currents were not inhibited by D₁-receptor activation (Fig. 4a, b, d), mimicking the subtype selectivity with which $\beta_2\gamma_2$ inhibits LVA channels (Fig. 1).

We tested further for a role for $\beta\gamma$, evaluating the importance of the α_{1H} II–III loop in receptor-coupled current inhibition by using the $\alpha_{1H}(G_{II-III})$ chimera. Consistent with the results shown in Fig. 3 was our observation that the $\alpha_{1H}(G_{II-III})$ chimera was not regulated by SKF 38393 (Fig. 4d). Finally, the C terminus of β -adrenergic receptor kinase-1 (β -ARK-CT) was used to provide a ‘sink’ for $\beta\gamma$ (ref. 25) and thereby prevented current inhibition. β -ARK-CT expression prevented the inhibition of α_{1H} current with SKF 38393 by about 50% (Fig. 4d). Taken together, these findings

provide support for a key function for β_2 -containing $\beta\gamma$ dimers in D₁-receptor-mediated current inhibition.

A general property of D₁ receptors is the activation of adenylyl cyclase²⁶. In addition to stimulation by α_s the type II and IV isoforms are stimulated by $\beta\gamma$ (ref. 27). To evaluate the relative importance of cAMP production in the mechanism of action of D₁ receptor-coupled current inhibition mediated by $\beta\gamma$, PKI (5–24)—a potent ($K_i = 2.3$ nM) and selective peptide inhibitor of cAMP-dependent protein kinase (PKA)—was included in the patch pipette. PKI failed to diminish the inhibitory effects of D₁-receptor stimulation even at a concentration of 1 μ M (Fig. 4d), which is consistent with a lack of effect of the active catalytic subunit of PKA on control α_{1H} currents or α_{1H} currents inhibited by $\beta_2\gamma_2$ (data not shown). Taken together, our findings do not support a previously suggested role for cAMP in the mechanism of D₁-receptor-coupled current inhibition²⁸ but instead highlight the importance of $\beta\gamma$.

The present study offers a molecular explanation of how T-type Ca^{2+} channel currents can be inhibited by G-protein-coupled receptors. We have uncovered a remarkable specificity in the interaction between $\beta\gamma$ dimers and LVA Ca^{2+} channels. The direct binding of only β_2 -containing dimers to the II–III loop of only the α_{1H} channel subtype confers current inhibition. Although inhibition is independent of voltage, β_2 -containing dimers do not change the surface expression of α_{1H} channel protein; they therefore control channel function rather than trafficking and/or turnover. The spatial constraints imposed by the membrane delimitation of $\beta\gamma$ dimers, and the extraordinary targeted specificity of their interaction as shown here, could provide the underpinnings for unique functional control in a dynamic signaling environment. □

Methods

Electrophysiology

Whole-cell currents were recorded from HEK-293 or H295R cells with the use of the voltage clamp method. Internal solution (in mM): 115 CsCl, 1 TBACl, 1 MgCl₂, 5 Mg-ATP, 1 Li-GTP, 20 HEPES pH 7.2 (adjusted with CsOH), 11 BAPTA; added CaCl₂ fixed free Ca^{2+} at 27 nM (0.9 mM) or 1 μ M (8.8 mM) with 2 μ M calmodulin. Bath solution (in mM): 127 tetraethylammonium chloride, 10 CaCl₂, 0.5 MgCl₂, 10 HEPES, 5 dextrose, 32 sucrose, pH 7.4 (adjusted with CsOH). Currents were filtered at 2.5 kHz and sampled at 12.5 kHz; except in prepulse experiments, leak subtraction was performed on line using scaled hyperpolarizing steps of one-fourth amplitude (P/N_4). Currents from $\beta\gamma$ -transfected cells were compared with control by analysis of variance with the post-hoc Dunnett test, where significance was taken as $P < 0.05$. Activation and inactivation gating were measured by comparing tail-current amplitudes elicited at -60 mV to I_{max} . For activation, depolarizations were in 5-mV increments (-60 to $+10$ mV; 10.4 ms) from a holding potential of -90 mV. For inactivation, cells were held at 15 potentials in 5-mV increments (-90 to -20 mV for 6 s) before depolarization to $+20$ mV and repolarization to -60 mV. Tail currents were fitted to a single exponential plus a constant by using the Chebyshev algorithm.

Preparation of chimaeric and mutant constructs

A truncated α_{1H} sequence (of AF 051946 Version 1) was extended by PCR to full length (AF 051946-modified). The II–III linker regions of α_{1H} (AF 051946-modified) and α_{1G} (AF 190860) were generated by PCR (α_{1H} , 5'-TCCTGTACAACGGCATGG-3' and 5'-CTCTCTAGATATCAGCGTCTCCACC-3'; α_{1G} , 5'-CCTGTACAATGGTATGGCCCTCC-3' and 5'-CTCTCTAGACAGTGGTGAACATCTTG-3'); subcloned into pUC18, resulting in pUC18- H_{II-III} and pUC18- G_{II-III} , and sequenced. The II–III linker of α_{1G} was isolated from pUC18- G_{II-III} by digestion with *Bsr*GI and *Pml*I and inserted into pUC18- H_{II-III} at the corresponding site to replace H_{II-III} , resulting in pUC18- $G_{II-III}-1$. The *Bsr*GI-*Eco*RV fragment containing the II–III linker of α_{1G} was isolated and subcloned back into pcDNA3- α_{1H} , resulting in construct H- G_{II-III} . The H_{II-III} linker was isolated and subcloned into pcDNA3.1(–) at the *Eco*RI-*Kpn*I site, resulting in the construct G- H_{II-III} .

Cell-surface biotinylation

α_{1H} cells were transfected with $\beta_2\gamma_2$ or empty vector with GFP as a transfection marker. GFP-positive cells were sorted with a FACS Vantage SE flow cytometer (Becton Dickinson) and plated in 35-mm dishes. Cells were washed with PBS and incubated twice for 15 min in biotinylation buffer (10 mM triethanolamine pH 9, 2 mM CaCl₂, 150 mM NaCl, 2 mg ml⁻¹ NHS-SS-biotin (Pierce)). Unreacted biotin was quenched with 100 mM glycine, and cells were washed and incubated for 1 h in lysis buffer (150 mM NaCl, 5 mM EDTA, 50 mM Tris-HCl pH 7.5, 1 μ M phenylmethylsulphonyl fluoride, 1% Triton X-100). After centrifugation for 2 min at 20,000g, lysate volume and protein concentration were normalized. Streptavidin-agarose beads (Pierce) were added to the lysates for 16 h. Beads were washed; bound proteins were eluted, resolved on an acrylamide gel and transferred to

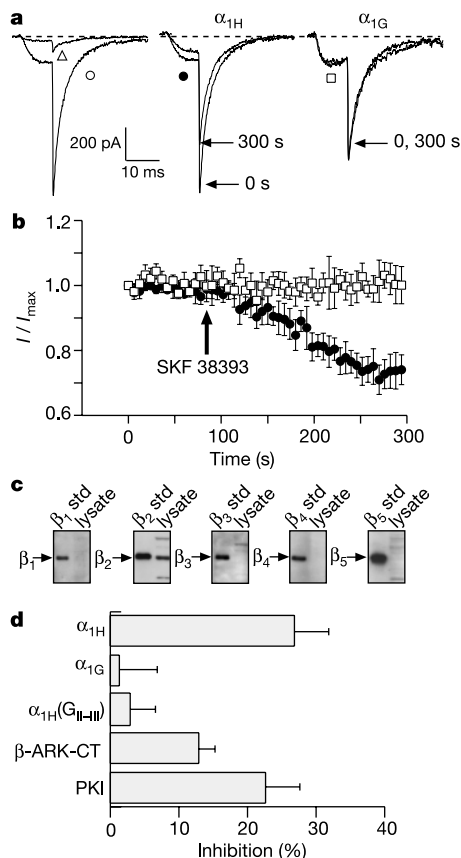


Figure 4 Effect of dopamine D₁-receptor activation on wild-type and chimaeric T-type Ca^{2+} channels. **a**, Left, sample tail currents from H295R cells before (open triangles) and after (open circles) heterologous expression of α_{1H} channels. Middle and right, sample currents from H295R cells transfected with α_{1H} (filled symbols) and α_{1G} (open symbols) at the indicated times. SKF 38393 (10 μ M), a selective D₁ agonist, was added at 60 s. **b**, Time course of SKF 38393-induced current inhibition expressed as a percentage of initial tail-current amplitude for α_{1H} (filled symbols, $n = 5$) and α_{1G} (open symbols, $n = 4$) channels. **c**, Immunodetection of β subunits in H295R cell lysates with β -subunit-selective antibodies; 30 ng of each recombinant β served to standardize blots. **d**, Histogram depicting percentage inhibition at 300 s as in **b** for α_{1H} ($n = 5$), α_{1G} ($n = 4$) or $\alpha_{1H}(G_{II-III})$ ($n = 4$) channels, and α_{1H} channels with co-expressed β -adrenergic kinase 1 C terminus (β -ARK-CT; $n = 5$) or PKA inhibitor peptide (1 μ M PKI) in the patch pipette solution ($n = 5$). Error bars are s.e.m.

poly(vinylidene difluoride) for immunodetection with anti- α_{1H} antibody and enhanced chemiluminescence. Films were scanned in a Molecular Dynamics Personal Densitometer and the integrated density of α_{1H} bands was quantified with ImageQuant 5.2 software.

$\beta\gamma$ binding experiments

Cytoplasmic II–III linkers of α_{1H} (3117–3696) and α_{1G} (2902–3762) were amplified by polymerase chain reaction, inserted into pGEX-2T (Pharmacia), and sequenced. The QXXER $\rightarrow$ AXAA mutation was introduced into the α_{1H} construct by using Stratagene's QuikChange XL Site-Directed Mutagenesis Kit. GST fusion proteins were expressed in BL21 cells (Stratagene) and purified with glutathione-conjugated agarose beads. $\beta\gamma$ subunits were purified from Genapol extracts of baculovirus-infected Sf9 cell membranes. Viruses expressed hexahistidine-tagged $G\alpha_{1H}$ and the indicated β and γ subunits. The extracts were applied to a Ni^{2+} -nitrilotriacetate column and $\beta\gamma$ subunits were eluted^{29,30}. Purified $\beta\gamma$ subunits were incubated for 3 h with α_{1H} II–III loop GST or α_{1G} II–III loop GST (200 nM) at 4 °C in binding buffer (20 mM HEPES pH 7.5, 3 mM $MgCl_2$, 150 mM NaCl, 1 mM 2-mercaptoethanol, 0.1% Genapol C-100 (Calbiochem)). The bead complex was pelleted, washed, boiled in sample loading buffer, separated by SDS–polyacrylamide-gel electrophoresis and immunoblotted with anti- β antisera (see below).

Antibodies

Anti-Flag antibody was from Sigma. Anti- β antibodies were from Santa Cruz (sc-261, 379, 380, 381, 382) except anti- β_5 (raised against the N terminus by Sigma for J.C.G.). Anti-GFP antibody was from Chemicon. Antibody against α_{1H} was a gift from Merck. Primary antibodies were detected by enhanced chemiluminescence with secondary antisera coupled to horseradish peroxidase.

Transfection

Stable cell lines (α_{1H} , α_{1G} , $\alpha_{1H}(G_{II-III})$ and $\alpha_{1G}(H_{II-III})$) were transiently transfected with plasmids expressing β (N-terminally Flag-tagged) and γ_2 subunits under the control of a cytomegalovirus promoter. Control cells were transfected with an equivalent amount of empty pCDNA3. Complementary DNAs were obtained from the Guthrie Research Institute. Cells were transfected in 35-mm culture dishes by $CaPO_4$ precipitation together with a plasmid (pGreenLantern; Gibco) for GFP in a ratio of 6:1 (test plasmids to GFP plasmid, by mass (μ g)). H295R cells were transfected with Effectene reagent (Qiagen) in accordance with the manufacturer's protocol. The β -ARK construct (W. J. Koch, Duke University, Durham, North Carolina) was subcloned in-frame with a myristic acid attachment signal in pGTM, a pCDNA3 derivative (E. A. Golemis, Fox Chase Cancer Center, Philadelphia, Pennsylvania).

Received 2 April; accepted 14 May 2003; doi:10.1038/nature01772.

- Perez-Reyes, E. Molecular physiology of low-voltage-activated T-type calcium channels. *Physiol. Rev.* **83**, 117–161 (2003).
- Lledo, P. M., Legendre, P., Israel, J. M. & Vincent, J. D. Dopamine inhibits two characterized voltage-dependent calcium currents in identified rat lactotroph cells. *Endocrinology* **127**, 990–1001 (1990).
- Lledo, P. M., Homburger, V., Bockaert, J. & Vincent, J. D. Differential G protein-mediated coupling of D2 dopamine receptors to K^+ and Ca^{2+} currents in rat anterior pituitary cells. *Neuron* **8**, 455–463 (1992).
- Marchetti, C., Carbone, E. & Lux, H. D. Effects of dopamine and noradrenaline on Ca channels of cultured sensory and sympathetic neurons of chick. *Pflügers Arch. Eur. J. Physiol.* **406**, 104–111 (1986).
- Williams, P. J., MacVicar, B. A. & Pittman, Q. J. Synaptic modulation by dopamine of calcium currents in rat pars intermedia. *J. Neurosci.* **10**, 757–763 (1990).
- Bean, B. P. Neurotransmitter inhibition of neuronal calcium currents by changes in channel voltage dependence. *Nature* **340**, 153–156 (1989).
- Ikeda, S. R. Voltage-dependent modulation of N-type calcium channels by G-protein $\beta\gamma$ subunits. *Nature* **380**, 255–258 (1996).
- Logothetis, D. E., Kurachi, Y., Galper, J., Neer, E. J. & Clapham, D. E. The $\beta\gamma$ subunits of GTP-binding proteins activate the muscarinic K^+ channel in heart. *Nature* **325**, 321–326 (1987).
- Herlitz, S. *et al.* Modulation of Ca^{2+} channels by G-protein $\beta\gamma$ subunits. *Nature* **380**, 258–262 (1996).
- De Waard, M. *et al.* Direct binding of G-protein $\beta\gamma$ complex to voltage-dependent calcium channels. *Nature* **385**, 446–450 (1997).
- Qin, N., Platano, D., Olcese, R., Stefani, E. & Birnbaumer, L. Direct interaction of $G\beta\gamma$ with a C-terminal $G\beta\gamma$ -binding domain of the Ca^{2+} channel α_1 subunit is responsible for channel inhibition by G protein-coupled receptors. *Proc. Natl Acad. Sci. USA* **94**, 8866–8871 (1997).
- Zhang, J. F., Ellinor, P. T., Aldrich, R. W. & Tsien, R. W. Multiple structural elements in voltage-dependent Ca^{2+} channels support their inhibition by G proteins. *Neuron* **17**, 991–1003 (1996).
- Zamponi, G. W., Bourinet, E., Nelson, D., Nargeot, J. & Snutch, T. P. Crosstalk between G proteins and protein kinase C mediated by the calcium channel α_1 subunit. *Nature* **385**, 442–446 (1997).
- Jarvis, S. E. & Zamponi, G. W. Interactions between presynaptic Ca^{2+} channels, cytoplasmic messengers and proteins of the synaptic vesicle release complex. *Trends Pharmacol. Sci.* **22**, 519–525 (2001).
- Catterall, W. A. Structure and regulation of voltage-gated calcium channels. *Annu. Rev. Cell Dev. Biol.* **16**, 521–555 (2000).
- Cooper, C. B. *et al.* Cross-talk between G-protein and protein kinase C modulation of N-type calcium channels is dependent on the G-protein β subunit isoform. *J. Biol. Chem.* **275**, 40777–40781 (2000).
- Downes, G. B. & Gautam, N. The G protein subunit gene families. *Genomics* **62**, 544–552 (1999).
- Lei, Q. *et al.* Activation and inhibition of G protein-coupled inwardly rectifying potassium ($Kir3$) channels by G protein $\beta\gamma$ subunits. *Proc. Natl Acad. Sci. USA* **97**, 9771–9776 (2000).
- Fishman, M. C. & Spector, I. Potassium current suppression by quinidine reveals additional calcium currents in neuroblastoma cells. *Proc. Natl Acad. Sci. USA* **78**, 5245–5249 (1981).
- Beguin, P. *et al.* Regulation of Ca^{2+} channel expression at the cell surface by the small G-protein $kir/$ *Gep*. *Nature* **411**, 701–706 (2001).
- Chen, J. *et al.* A region of adenylyl cyclase 2 critical for regulation by G protein $\beta\gamma$ subunits. *Science* **268**, 1166–1169 (1995).

- Safa, P., Boulter, J. & Hales, T. G. Functional properties of Cav1.3 (α_{1D}) L-type Ca^{2+} channel splice variants expressed by rat brain and neuroendocrine GH3 cells. *J. Biol. Chem.* **276**, 38727–38737 (2001).
- Schrier, A. D., Wang, H., Talley, E. M., Perez-Reyes, E. & Barrett, P. Q. α_{1H} T-type Ca^{2+} channel is the predominant subtype expressed in bovine and rat zona glomerulosa. *Am. J. Physiol. Cell Physiol.* **280**, C265–C272 (2001).
- Bird, I. M. *et al.* Human NCI-H295 adrenocortical carcinoma cells: A model for angiotensin-II-responsive aldosterone secretion. *Endocrinology* **133**, 1555–1561 (1993).
- Koch, W. J., Hawes, B. E., Inglese, J., Luttrell, L. M. & Lefkowitz, R. J. Cellular expression of the carboxyl terminus of a G protein-coupled receptor kinase attenuates G $\beta\gamma$ -mediated signaling. *J. Biol. Chem.* **269**, 6193–6197 (1994).
- Missale, C., Nash, S. R., Robinson, S. W., Jaber, M. & Caron, M. G. Dopamine receptors: from structure to function. *Physiol. Rev.* **78**, 189–225 (1998).
- Sunahara, R. K., Dessauer, C. W. & Gilman, A. G. Complexity and diversity of mammalian adenylyl cyclases. *Annu. Rev. Pharmacol. Toxicol.* **36**, 461–480 (1996).
- Drolet, P. *et al.* Inhibition of the T-type Ca^{2+} current by the dopamine D1 receptor in rat adrenal glomerulosa cells: Requirement of the combined action of the G $\beta\gamma$ protein subunit and cyclic adenosine 3',5'-monophosphate. *Mol. Endocrinol.* **11**, 503–514 (1997).
- Kozasa, T. in *G Proteins, Techniques for Analysis* (ed. Manning, D. R.) 23–38 (CRC, Boca Raton, Florida, 1999).
- Graber, S. G., Lindorfer, M. A. & Garrison, J. C. Expression and purification of G-protein $\beta\gamma$ subunits using a baculovirus expression system. *J. Methods Neurosci.* **29**, 207–226 (1996).

Acknowledgements We thank E. Perez-Reyes for α_{1H} and α_{1G} clones, the University of Virginia Flow-Cytometry Core for FACS analysis, and G. MacCleery for help with $\beta\gamma$ purification. This study was supported by grants (to P.Q.B. and J.C.G.) from the NIH. J.T.W. was supported by predoctoral fellowships from the University of Virginia Cardiovascular Research Center and the American Heart Association.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.Q.B. (pqb4b@virginia.edu).

HIV-1 Nef intersects the macrophage CD40L signalling pathway to promote resting-cell infection

Simon Swingler, Beda Brichacek, Jean-Marc Jacque, Catherine Ulich, Jin Zhou & Mario Stevenson

Program in Molecular Medicine, University of Massachusetts Medical School, 373 Plantation Street, Worcester, Massachusetts 01605, USA

All primate lentiviruses (HIV-1, HIV-2, SIV) encode Nef proteins, which are important for viral replication and pathogenicity *in vivo*^{1–3}. It is not known how Nef regulates these processes. It has been suggested that Nef protects infected cells from apoptosis and recognition by cytotoxic T lymphocytes^{4–6}. Other studies suggest that Nef influences the activation state of the infected cell, thereby enhancing the ability of that cell to support viral replication^{7–10}. Here we show that macrophages that express Nef or are stimulated through the CD40 receptor release a paracrine factor that renders T lymphocytes permissive to HIV-1 infection. This activity requires the upregulation of B-cell receptors involved in the alternative pathway of T-lymphocyte stimulation. T lymphocytes stimulated through this pathway become susceptible to viral infection without progressing through the cell cycle. We identify two proteins, soluble CD23 and soluble ICAM, that are induced from macrophages by Nef and CD40L, and which mediate their effects on lymphocyte permissivity. Our results reveal a mechanism by which Nef expands the cellular reservoir of HIV-1 by permitting the infection of resting T lymphocytes.

We have previously shown that human immunodeficiency virus (HIV)-1 Nef induces the release of CC- chemokines from infected macrophages, and we have proposed that Nef may promote the

recruitment of substrate T lymphocytes to sites of infection¹¹. We initiated the current study when we observed that lymphocyte/macrophage co-cultures supported the replication of wild-type HIV-1 more efficiently than did Nef-deleted virus, and that T lymphocytes within these co-cultures did not require exogenous stimulation to support viral replication. The ability of peripheral blood lymphocytes (PBLs) to support virus (HIV-1_{LAI}) replication was examined after their incubation with macrophages harbouring a wild-type (HIV-1_{SFI} WT) or Δ Nef virus (HIV-1_{SFI} Δ Nef) (Fig. 1a). The level of virus production in macrophages infected with HIV-1_{SFI} Δ Nef approximated (twofold difference) that in wild-type HIV-1_{SFI}-infected macrophages (Fig. 1b, left panel). Despite this, infectious HIV-1_{LAI} (determined by virus titration on MT-4 cells) was detected only when autologous PBLs were co-cultured with wild-type HIV-1_{SFI}-infected macrophages and not HIV-1_{SFI} Δ Nef-infected or uninfected macrophages (Fig. 1b, right panel). Similarly, lymphocytes cultured in the absence of infected macrophages remained refractory to HIV-1_{LAI} replication (not shown). Thus, in infected macrophages, Nef induces a stimulus that renders

lymphocytes permissive to infection. Nef activates the production of the CC- chemokines MIP-1 α and MIP-1 β ¹¹. These chemokines are also regulated, in a NF- κ B-dependent manner, through the co-stimulatory receptor CD40 (ref. 12). We observed that Nef, in the context of viral replication and when expressed from an adenovirus vector, induced NF- κ B activation (not shown), raising the possibility that signals generated by Nef and CD40 intersect.

To examine the relationship between Nef and CD40 signalling, T lymphocytes were incubated with supernatants from macrophages activated by CD40L and examined for permissivity to infection. Whereas lymphocytes exposed to supernatants of untreated macrophages were completely refractory to HIV-1 replication, those exposed to culture supernatants from CD40L-stimulated macrophages were permissive to HIV-1 infection (Fig. 1c). This property of CD40L-stimulated macrophages was identical for Nef-expressing macrophages in that it required neither cell contact nor exogenous activation of substrate lymphocytes. We next examined whether the ability of CD40L and Nef to induce lymphocyte permissivity was dependent on NF- κ B. We used an

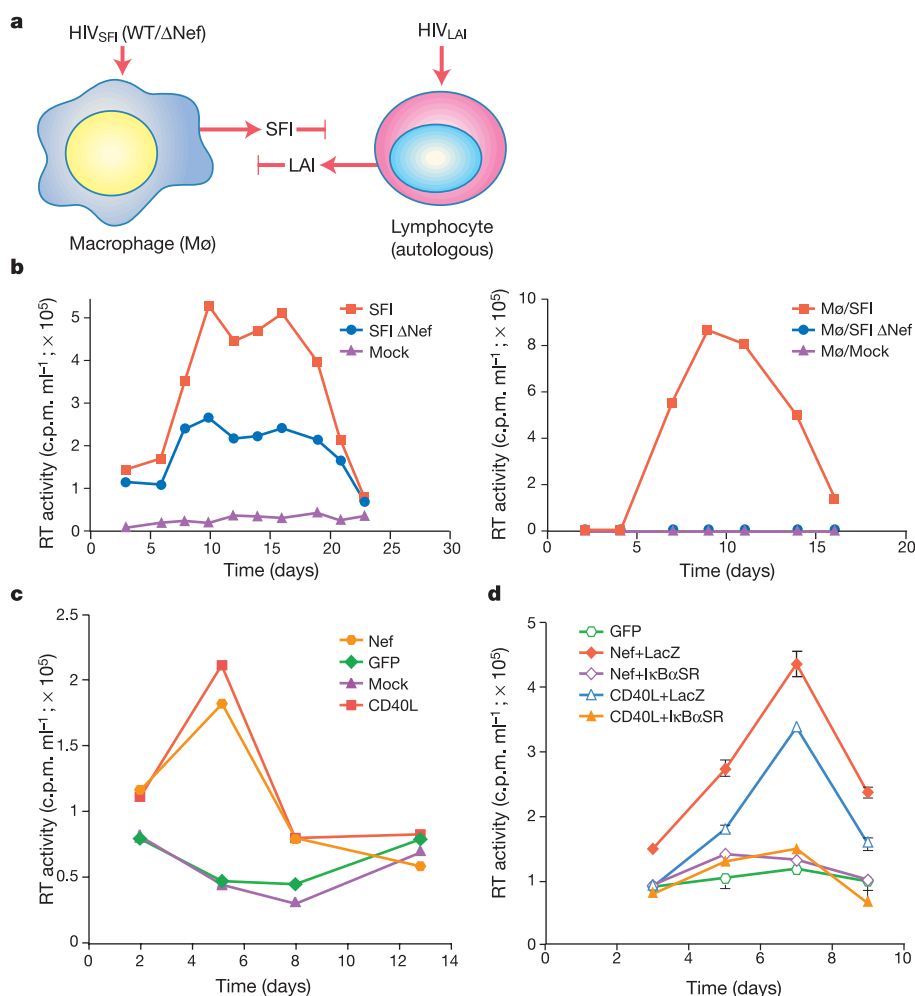


Figure 1 Macrophages expressing Nef render lymphocytes permissive to HIV-1 replication. **a**, Assay for permissivity. The tropism HIV-1_{SFI} and HIV-1_{LAI} is restricted to macrophages and T cells, respectively. HIV-1_{LAI} production (detected by titration on MT4 cells) provides an indication that those lymphocytes were rendered permissive for viral replication by infected macrophages. **b**, Wild-type but not Nef-deleted viruses induce lymphocyte permissivity. Virus production from HIV-1_{SFI} WT- and Δ Nef-infected macrophages in the absence of lymphocyte culture (left panel); MT4 titration of HIV-1_{LAI} in supernatants of lymphocytes co-cultured with HIV-1_{SFI} WT-infected, HIV-1_{SFI} Δ Nef-

infected and mock-infected macrophages (right panel). **c**, CD40 stimulation induces lymphocyte permissivity. Viral replication was examined in resting lymphocytes exposed to supernatants from Nef- and GFP-expressing macrophages, or from CD40L-stimulated macrophages. **d**, Inhibition of Nef- and CD40-induced permissivity by an IkB α 'super-repressor'. Macrophages were subject to CD40L stimulation/adenovirus infections as indicated, and supernatants were examined for induction of permissivity. RT, reverse transcriptase.

adenovirus expressing a 'super-repressor' variant of I κ B α (I κ B α SR)¹³ to inhibit NF- κ B activation. In the presence of I κ B α SR, the ability of Nef to induce lymphocyte permissivity was inhibited (Fig. 1d). Co-infection with an adenoviral vector expressing a control protein (LacZ) had no effect on Nef activity (Fig. 1d). The I κ B α SR adenovirus, but not the control adenovirus, also inhibited the ability of CD40L-stimulated macrophages to induce lymphocyte permissivity (Fig. 1d). Macrophage function, as determined from the levels of macrophage-colony stimulating factor (MCSF) released into culture supernatants, was not impaired following transduction with the adenovirus vectors (not shown). Collectively, these results indicate that Nef induces lymphocyte permissivity to HIV-1 infection by intersecting a pathway that is regulated by the CD40 receptor, and that CD40 activation can functionally substitute for Nef.

Because quiescent (G0) T lymphocytes are completely refractory to HIV-1 replication *in vitro*, we expected Nef and CD40L to induce permissivity by promoting lymphocyte activation. We reasoned that an analysis of specific CD4⁺ lymphocyte subsets stimulated by Nef or CD40L would assist in identifying the mechanism through which lymphocyte permissivity is effected. However, our experiments revealed that T lymphocytes, when separated from B lymphocytes, did not become permissive to HIV-1 infection on exposure to supernatants of Nef-expressing or CD40L-stimulated macrophages (Fig. 2). T-cell permissivity could be induced by B lymphocytes that were pretreated with supernatants from Nef-expressing or CD40L-stimulated macrophages (Fig. 2a). As supernatants from these

pretreated B lymphocytes were inactive (Fig. 2a), this indicated a role for accessory molecules on B lymphocytes in the induction of T-lymphocyte permissivity by Nef and CD40L. Immunophenotyping revealed that B lymphocytes exposed to supernatants from Nef-expressing macrophages exhibited increased expression of the receptors CD22, CD54 and CD58—a pattern that was identical for B lymphocytes exposed to supernatants from CD40L-stimulated macrophages (Fig. 2b). Since antibody-blocking experiments indicated that these receptors were involved in the induction of permissivity by Nef and CD40L (not shown), we examined whether direct stimulation of the ligands for these receptors on T lymphocytes would render them permissive to HIV-1 infection.

Purified T lymphocytes were stimulated with antibodies to CD45 RO/RA, CD11a and CD2 (the T-cell ligands for CD22, CD54 and CD58, respectively), then infected with a recombinant HIV-1 variant (HIV-1_{HSA})¹⁴. Expression of heat-stable antigen (HSA) indicates completion of steps in the viral life cycle up to, and including, *de novo* viral gene expression. CD2 and CD45 ligation on T lymphocytes was sufficient to render those cells highly susceptible to HIV-1 infection at rates approaching those obtained for T lymphocytes receiving a potent—PHA (phytohaemagglutinin)/interleukin-2 (IL-2)—activation stimulus (Fig. 2c). By contrast, other antibodies, such as CTLA-4, did not promote lymphocyte permissivity, nor did the cytokines IL-1 or IL-16 (not shown). The induction of lymphocyte permissivity minimally required a CD2 stimulus while the RO and RA isoforms of CD45 were interchangeable, and the CD11a stimulus augmented the

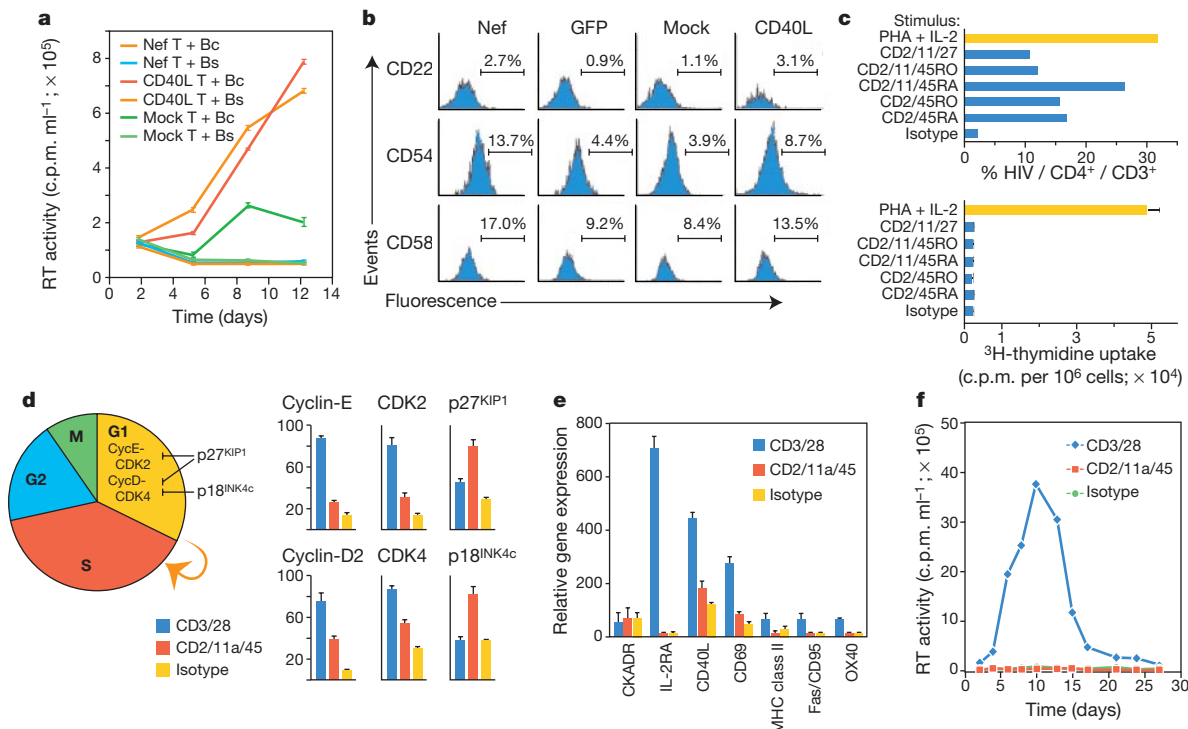


Figure 2 Nef and CD40 promote resting T-lymphocyte infection by regulating co-stimulatory receptors on B lymphocytes. **a**, Nef- and CD40-mediated permissivity requires the presence of B lymphocytes. B lymphocytes were incubated with supernatants from Nef-expressing or CD40L-stimulated macrophages. After 72 h, the B cells (Bc) or B-cell supernatants (Bs) were added to autologous lymphocytes, which were then examined for susceptibility to HIV-1 infection. **b**, Nef and CD40 upregulate similar receptors on B cells. B lymphocytes were incubated with supernatants from Nef- or GFP-expressing or CD40L-treated macrophages, and expression of CD22, CD54 and CD58 on CD19⁺ B cells was determined by FACS. **c**, B-lymphocyte signals induce cell-cycle-independent infection of T lymphocytes. Following ligation of the indicated

receptors, resting CD3⁺ T lymphocytes were infected with HIV-1_{HSA} and the level of infection and cell-cycle status determined by FACS (upper panel) and by thymidine incorporation (lower panel), respectively. Rabbit F(ab)₂-cross-linked mouse IgG (isotype) is shown as a control. **d**, Expression profiles of cell-cycle regulatory genes in antibody-stimulated lymphocytes. Purified CD3⁺ lymphocytes were stimulated as in **c** and transcript levels for the indicated genes determined by RNase protection assay. **e**, Transcript profiles of cellular activation markers. **f**, Viral replication profiles in antibody-stimulated T lymphocytes. CycE, Cyclin-E; CycD, Cyclin-D; MHC, major histocompatibility complex.

CD2/CD45 stimulus (Fig. 2c). Contrary to our expectations, T lymphocytes rendered permissive to infection by the CD2/CD45 stimulus had not progressed through the cell cycle, on the basis of the level of thymidine uptake (Fig. 2c). Transition beyond G1 requires the cooperative action of Cyclin-D and Cyclin-E with their catalytic kinase partners CDK4 (cyclin-dependent kinase 4) and CDK2, which in turn are regulated by p18^{INK4c} and p27^{KIP-1} (Fig. 2d). Although activated (CD3/CD28) lymphocytes exhibited higher levels of Cyclin/CDK transcripts relative to CDK-inhibitor transcripts, this relationship was inverted in CD2/CD11a/CD45-stimulated cells (Fig. 2d), indicating that this stimulus was insufficient to overcome blocks that prevent G1-to-S transition. CD2/CD11a/CD45-stimulated T lymphocytes also lacked transcripts of genes that are expressed in activated cells (Fig. 2e). Although CD2/11a/45 stimulation of CD4⁺ T lymphocytes allowed efficient viral infection and *de novo* expression of viral proteins, as evidenced by HSA expression (Fig. 2c), this stimulus did not allow the efficient release of virions (Fig. 2f). Thus, the CD2/CD11a/CD45 stimulus was sufficient to remove blocks leading up to the expression of viral proteins but not virus release.

As ligation of CD2/CD45 on T lymphocytes only partially recapitulated the activities of Nef and CD40, we did further immunophenotyping of B cells to identify receptors that might promote a productive infection. FACS analysis further revealed the presence of distinct subpopulations of CD22/CD58-positive B cells that had differentially upregulated the CD80/CD86 receptors in response to Nef and CD40L (Fig. 3a). The induction of T-lympho-

cyte permissivity by Nef and CD40 was strongly inhibited when the B lymphocytes were pretreated with an anti-CD80 antibody (Fig. 3b). By contrast, antibody to CD86 partially impaired CD40L activity and had no effect on the activity of Nef (Fig. 3b). CD80/CD86 interaction with their T-cell ligand CD28 has a central role in T-lymphocyte co-stimulation^{15,16}. Ligating CD28 on T lymphocytes augmented the CD2/CD11a/CD45 stimulus by inducing T lymphocytes to enter cell cycle (Fig. 3c), and recapitulated a productive viral infection (Fig. 3d).

We next undertook a characterization of the factor(s) involved in the induction of permissivity by Nef and CD40L. We took advantage of our observation (Figs 1 and 2) that the induction of T-lymphocyte permissivity was recapitulated by CD40L stimulation and required the involvement of B-cell receptors. Therefore, we focused our analysis on soluble proteins implicated in CD40-mediated effector functions. We identified two proteins—the soluble form of the intercellular adhesion molecule ICAM-1 (sICAM) and the soluble form of the coactivation molecule CD23 (sCD23)—that were elevated in the supernatants of macrophages infected with wild-type HIV-1 (Fig. 4a). By contrast, sCD23 and sICAM levels in HIV-1 Δ Nef-infected macrophage supernatants were similar to those in uninfected cultures (Fig. 4a). These differences in sCD23 and sICAM were apparent despite equivalent levels of wild-type and Δ Nef HIV-1 replication. MCSF, which is induced by HIV-1 replication¹⁷ in a Nef-independent manner¹¹, was present at comparable levels in wild-type and Δ Nef HIV-1-infected cultures (Fig. 4a). sCD23 and sICAM were also induced from macrophages expressing Nef alone or those stimulated with CD40L (Fig. 4a, right panels). Immunodepletion of sCD23 and sICAM from supernatants of Nef-expressing or CD40L-stimulated macrophages abrogated the ability of the supernatants to induce T-lymphocyte permissivity (Fig. 4b). Purified B cells incubated with sCD23 at a concentration similar to that induced in wild-type infected macrophages and Nef-expressing macrophages exhibited increased expression of CD22 and CD58, and to a lesser extent CD80 and CD86 (Fig. 4c). By comparison, sICAM induced the expression of CD80/CD86, and to a lesser extent CD22/CD58.

We next examined the impact of these proteins on the susceptibility of T lymphocytes to infection. When T lymphocytes were incubated with sCD23-treated B cells and subsequently challenged with HIV-1, infection (on the basis of HSA expression) was almost completely restricted to non-cycling (Ki67⁻) cells (Fig. 4d). By comparison, HIV-1 infection in T lymphocytes exposed to sICAM-treated B cells occurred in resting and cycling cells (Fig. 4d). The low level of resting-cell infection imparted by the isotype or CD3/CD28 stimulus (Fig. 4d, left panel) was probably due to the underlying presence of CD22/CD58-positive B cells in these cultures (Fig. 4c). Consistent with the requirement for B cells in Nef- and CD40-mediated permissivity, direct treatment of T lymphocytes with sCD23 or sICAM did not promote permissivity. Similarly, T cells exposed to sCD23 and sICAM and then incubated with untreated B cells did not support HIV-1 infection (not shown). Whereas the sICAM stimulus was sufficient to promote the entry of T lymphocytes into the cell cycle, as evidenced by an increase in the levels of Ki67⁺ cells, the sCD23 stimulus was not (not shown). This suggests a threshold in the B-cell stimulus that promotes the entry of T lymphocytes into the cell cycle. T lymphocytes incubated with sICAM-treated B cells also supported an efficient spreading infection (Fig. 4d, right panel). By contrast, low levels of viral replication were evident in T-lymphocyte cultures incubated with sCD23-treated B cells (Fig. 4d, right panel). This restricted pattern of infection in sCD23-stimulated cultures (Fig. 4d) parallels that invoked by the CD22/CD58 stimulus (Fig. 2). When sCD23 and sICAM were added in combination, their effect on lymphocyte permissivity was not additive (not shown). Collectively, these results suggest that the activities of sCD23 and sICAM are not completely redundant, and that although both proteins can, to differing

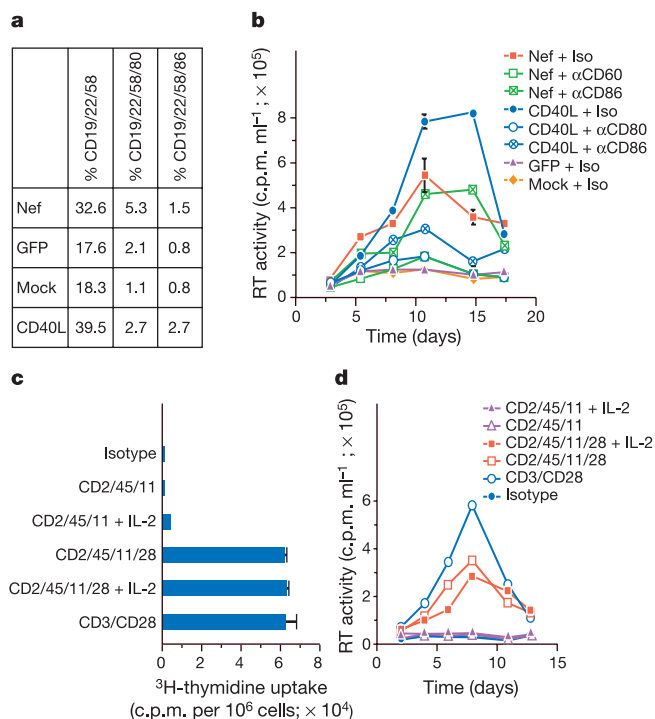


Figure 3 Identification of Nef and CD40-regulated B-lymphocyte receptors that overcome blocks to virus production. **a**, Nef and CD40 upregulate CD80/CD86 expression on B lymphocytes. B lymphocytes were immunophenotyped following incubation with supernatants from Nef- or GFP-expressing on CD40L-stimulated macrophages. **b**, Antibody to CD80 blocks induction of permissivity by Nef and CD40. Supernatants of Nef-expressing and CD40L-stimulated macrophages were incubated with autologous, resting lymphocytes in the presence of α CD80, α CD86 or isotype (Iso) antibodies, then examined for the ability to support HIV-1 replication. **c**, The additional CD80 stimulus induces cell-cycle progression. The isotype control comprises Rabbit F(ab)₂-cross-linked murine IgG. The additional CD80 stimulus reconstitutes a productive infection.

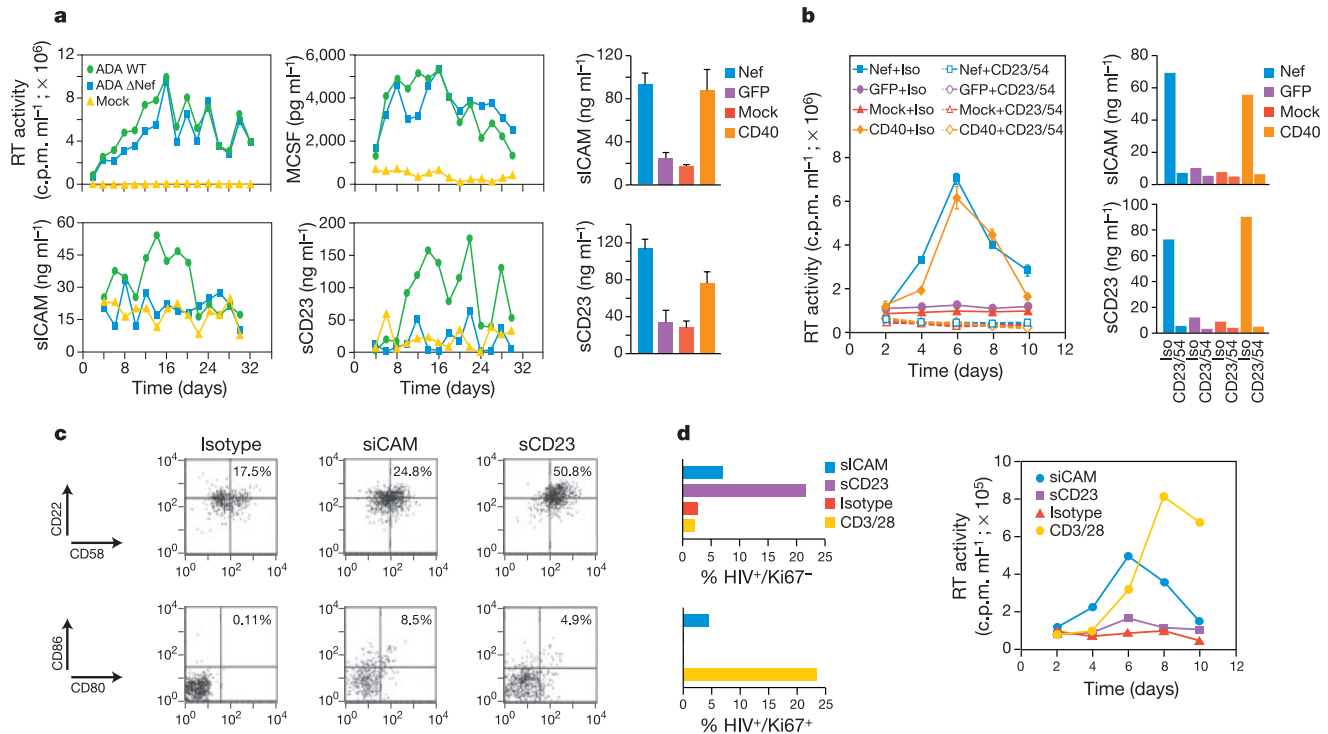


Figure 4 Induction of lymphocyte permissivity by Nef and CD40 is mediated by sCD23 and sICAM. **a**, Nef and CD40 induce sCD23 and sICAM release from macrophages. Replication profiles for wild-type and Δ Nef viruses are indicated along with levels of sICAM, sCD23 and MCSF in wild-type-infected, Δ Nef-infected and uninfected macrophage cultures. Results are representative of three independent donors. Right panels, sICAM and sCD23 levels in supernatants of macrophages 18 h after transduction with a Nef adenovirus vector or after CD40L stimulation. Levels in adeno-GFP-infected and mock-infected cultures are shown for comparison. **b**, Effects of sCD23 and sICAM immunodepletion on induction of permissivity by Nef or CD40L. Left panel, levels of viral replication. Right panels, levels of sICAM and sCD23 in immunodepleted macrophage

supernatants. Supernatants from adeno-GFP-transduced and mock-infected cultures serve as controls. **c**, sICAM and sCD23 upregulate receptors on B lymphocytes. Purified B lymphocytes were immunophenotyped after incubation with recombinant sICAM (50 ng ml⁻¹) or sCD23 (75 ng ml⁻¹). Cells were gated on CD19⁺ and CD58⁺ B cells for CD22/CD58 and CD80/CD86 analysis, respectively. **d**, sCD23 and sICAM promote T-lymphocyte permissivity. T lymphocytes were incubated with sICAM, sCD23 or isotype-treated B cells then infected with HIV-1_{HSA}. Infected resting (HIV⁺_{HSA}/Ki67⁻) and cycling (HIV-1_{HSA}/Ki67⁺) T lymphocytes were identified by FACS (left panels). Levels of virus replication in T-lymphocyte cultures stimulated with sCD23, sICAM or isotype-treated B cells (right panel).

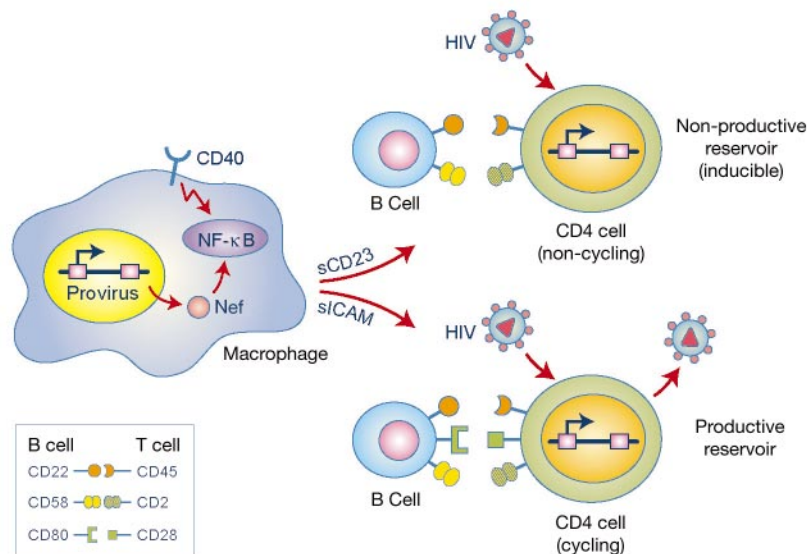


Figure 5 Mechanistic model for induction of lymphocyte permissivity by Nef and CD40. Physiological stimulation of CD40 in macrophages leads to NF- κ B activation. This signalling pathway is intersected by HIV-1 Nef. In macrophages, activation of CD40 or expression of Nef induces the release of sCD23 and sICAM, which upregulate the expression of co-stimulatory receptors on B lymphocytes. These, in turn, interact with their corresponding ligands on T lymphocytes, rendering them permissive to HIV-1

infection. The types of receptor that are upregulated on B cells dictate the outcome of the infection. The induction of CD22 and CD58 on B cells is mediated primarily by sCD23. The action of these receptors on T cells does not lead to T-cell proliferation, but is sufficient to permit virus entry and *de novo* expression of viral proteins, but not virion release. Induction of CD80 on B cells, as mediated primarily by sICAM, provides signals that promote entry of T cells into the cell cycle, thereby allowing the productive infection of these cells.

degrees, promote resting-cell infection, the productive infection of cycling cells requires sICAM.

The results presented here support a model (Fig. 5) in which HIV-1 Nef intersects the CD40 signalling pathway in macrophages to promote the release of sCD23 and sICAM. These proteins, in turn, promote interactions between B cells and T cells that render the T cells permissive to HIV-1 infection. An unexpected finding was that this activity of Nef permitted the efficient infection of resting cells by HIV-1. Experimental systems using lymphoid-organ cultures have indicated that the infection of resting lymphocytes is independent of the infection frequency of macrophages¹⁸. However, the stimulation of B cells by adhesion molecules expressed on stromal cells within peripheral lymphoid tissues¹⁹ might contribute to an underlying level of resting-cell infection. Future studies should determine whether infection of resting cells in lymphoid-organ cultures requires the B cell–T cell interplay identified in this study.

Our results could explain several enigmas regarding HIV-1 biology. As replication of HIV-1 *in vitro* is restricted to T lymphocytes that are in cycle^{20–23}, many studies have focused on the possibility that Nef enhances viral replication by promoting T-lymphocyte activation^{24,25}. However, this first requires the *de novo* expression of Nef in a cell that is refractory to infection. Although the expression of Nef from unintegrated DNA may occur in resting cells²³, our results invoke an alternative mechanism in which Nef activates signals on antigen-presenting cells that, in turn, allow T lymphocytes to exit the G0 stage of the cell cycle, thereby rendering them susceptible to infection (Fig. 5). Our results further reveal that the central component of the Nef stimulus (CD58) was mediated not by classical (CD3) pathways of T-cell stimulation, but through the alternative pathway (CD2) of T-lymphocyte stimulation²⁶. The CD2–CD58 interaction, which facilitates initial contact between the T lymphocyte and the antigen-presenting cell before specific antigen recognition, has not generally been viewed as one sufficient to drive T-lymphocyte proliferation²⁶. Our studies show that this stimulus, although insufficient to promote cell-cycle progression, was sufficient to permit HIV-1 infection. These results may provide insight into the mechanism through which the reservoir of resting T lymphocytes is established *in vivo*. The infection of non-cycling CD4⁺ T lymphocytes, characterized by the expression of viral proteins, has been demonstrated in the setting of acute simian immunodeficiency virus (SIV) and HIV-1 infection *in vivo*²⁷. We propose that Nef promotes conditions for resting-cell infection, thereby extending the cellular reservoir of HIV-1. Although the full significance of the resting-T-cell reservoir in HIV-1 pathogenesis is unclear, studies suggest that these cells are more stable than infected cycling cells²⁷ and, as such, may serve as a more persistent viral reservoir. □

Methods

Primary lymphocytes and macrophages

Lymphocytes and monocytes were obtained by leukapheresis from normal donors seronegative for HIV-1 and hepatitis B. Monocytes were further separated by counter-current centrifugal elutriation as detailed elsewhere²⁸. T and B lymphocytes were further purified by negative selection with antibody-coupled para-magnetic beads (Dynal). Cell purity was determined by flow-cytometric staining with fluorochrome antibodies (Pharmingen) to CD3 (T cell), CD19 (B cell) and CD45 (leukocyte). Elutriated monocytes were cultured for 2 days in medium containing MCSF (R&D Systems) and for a further 5 days in medium lacking MCSF, and then used for virus infections.

Transcript profiles

For analysis of activation markers in lymphocytes stimulated with co-receptor antibodies, total RNA was isolated 72 h after antibody stimulation. RNA was reverse transcribed in the presence of 250 µCi α-³²P-dATP and α-³²P-dCTP using gene-specific primers according to the manufacturer's protocol (Super Array). Filter arrays were hybridized and washed according to the manufacturer's instructions. Filters were visualized on a Phosphor-Imager (Molecular Dynamics, ImageQuant) and hybridization intensity quantified by Image Master Software (Amersham). For analysis of Cyclin-CDK and CDK-inhibitor transcripts in antibody-stimulated lymphocytes, total RNA was isolated 72 h after antibody stimulation. RNase protection assays were performed using ³²P-labelled riboprobes generated from assembled templates as detailed by the manufacturer

(Pharmingen). Products were resolved on 5% acrylamide–urea gels, visualized on a Phosphor-Imager and quantified using ImageQuant.

Viruses

HIV-1 stocks were obtained after transfection of HeLa or 293T cells with 25 µg proviral DNA, normalized on the basis of reverse-transcription activity and used directly for cell infections. HIV-1_{HSA}¹⁴ expresses mouse HSA in place of Vpr, allowing infected cells to be identified by flow cytometry. HSA expression indicates completion of steps in the viral life cycle up to, and including, *de novo* viral gene expression. Adenoviral vectors for Nef and GFP (green fluorescent protein) were prepared as described previously¹¹. Adenoviral vectors for IkBαZ have previously been described¹³. These vectors were amplified and titred in 293T cells.

Lymphocyte assays

For the induction of T-lymphocyte permissivity by specific antibodies, T lymphocytes purified by negative selection were incubated with antigen-specific or isotype-control murine monoclonal antibodies (Pharmingen), (2 µg ml⁻¹) and cross-linked with rabbit F(ab')₂ anti-mouse antibody (ICN Biomedicals) at 4 µg ml⁻¹. DNA synthesis (³H-thymidine incorporation) and susceptibility to HIV-1 infection was determined 96 h after antibody stimulation. Recombinant soluble CD40L (10 µg ml⁻¹) was mixed with an enhancing reagent (Alexis Corporation) at 1 µg ml⁻¹ to promote trimer formation. For Ki67 and HSA analysis, cells were first stained for the surface antigens CD3–CyChrome and HSA–phycoerythrin, then permeabilized with 1% paraformaldehyde/1% Tween, incubated with Ki67–FITC (fluorescein isothiocyanate) fluorochrome and analysed by FACS. Isotype-matched non-specific immunoglobulin-γ (IgG) served as a staining control.

Permissivity assay

Monocyte-derived macrophages were infected with wild type (HIV-1_{SF1}WT) or Nef-deleted (HIV-1_{SF1}ΔNef) viruses. These viruses also lacked a functional envelope gene and were thus capable of a single round of infection when pseudo-typed with envelope glycoproteins of vesicular stomatitis virus (VSV). Fifteen days after HIV-1_{SF1} infection, macrophages were co-cultured with autologous, resting PBLs that had been preinfected with the T-cell tropic virus HIV-1_{LAI}. As resting PBLs are refractory to productive viral infection, replication of HIV-1_{LAI} offers an indication that T lymphocytes are rendered susceptible to infection by HIV-1_{SF1}-infected macrophages. One week after initiation of the co-cultures, culture supernatants were removed and titred on the T-cell line MT4. Although these supernatants contain a mixture of HIV-1_{SF1} and HIV-1_{LAI}, only HIV-1_{LAI} virions will be scored in this assay.

sCD23/sICAM immunodepletion

Macrophage supernatants were incubated with 5 µg ml⁻¹ anti-CD23 and 5 µg ml⁻¹ anti-CD54 (Pharmingen), and 40 µl protein A/G beads (Santa Cruz Biotech) for 16 h at 4 °C. Residual antibody was removed by two successive treatments of culture supernatants with 40 µl protein A/G beads (2 h, 4 °C). This treatment was sufficient to remove the antibodies by approximately 99% as determined by IgG ELISA (enzyme-linked immunosorbent assay). As controls, macrophage supernatants were treated as above with isotype-matched IgG at the same concentrations. Immunodepleted supernatants were sterilized by filtration (0.22 µm) and analysed for sCD23, sICAM and MCSF by ELISA. Recombinant human sCD23 and the sCD23 ELISA were obtained from Pharmingen. Recombinant human sICAM-1 was obtained from BenderMed Systems; ELISAs for sICAM and MCSF were from R&D Systems.

Received 4 February; accepted 8 April 2003; doi:10.1038/nature01749.

- Kirchhoff, F., Greenough, T. C., Brettler, D. B., Sullivan, J. L. & Desrosiers, R. C. Brief report: absence of intact *nef* sequences in a long-term survivor with nonprogressive HIV-1 infection. *N. Engl. J. Med.* **332**, 228–232 (1995).
- Deacon, N. J. *et al.* Genomic structure of an attenuated quasi species of HIV-1 from a blood transfusion donor and recipients. *Science* **270**, 988–991 (1995).
- Kestler, H. W. *et al.* Importance of the *nef* gene for maintenance of high virus loads and for development of AIDS. *Cell* **65**, 651–662 (1991).
- Johnson, W. E. & Desrosiers, R. C. Viral persistence: HIV's strategies of immune system evasion. *Annu. Rev. Med.* **53**, 499–518 (2002).
- Collins, K. L. & Baltimore, D. HIV's evasion of the cellular immune response. *Immunol. Rev.* **168**, 65–74 (1999).
- Geleziunas, R., Xu, W., Takeda, K., Ichijo, H. & Greene, W. C. HIV-1 Nef inhibits ASK1-dependent death signalling providing a potential mechanism for protecting the infected host cell. *Nature* **410**, 834–838 (2001).
- Saksela, K., Cheng, G. & Baltimore, D. Proline-rich (PxxP) motifs in HIV-1 Nef bind to SH3 domains of a subset of Src kinases and are required for the enhanced growth of Nef + viruses but not for down-regulation of CD4. *EMBO J.* **14**, 484–491 (1995).
- Fackler, O. T., Luo, W., Geyer, M., Alberts, A. S. & Peterlin, B. M. Activation of Vav by Nef induces cytoskeletal rearrangements and downstream effector functions. *Mol. Cell* **3**, 729–739 (1999).
- Fackler, O. T. & Baur, A. S. Live and let die: Nef functions beyond HIV replication. *Immunity* **16**, 493–497 (2002).
- Simmons, A., Aluvihare, V. & McMichael, A. Nef triggers a transcriptional program in T cells imitating single-signal T cell activation and inducing HIV virulence mediators. *Immunity* **14**, 763–777 (2001).
- Swingler, S. *et al.* HIV-1 Nef mediates lymphocyte chemotaxis and activation by infected macrophages. *Nature Med.* **5**, 997–1003 (1999).
- van Kooten, C. & Banchereau, J. CD40–CD40 ligand. *J. Leukoc. Biol.* **67**, 2–17 (2000).
- Wang, C. Y., Cusack, J. C. Jr, Liu, R. & Baldwin, A. S. Jr Control of inducible chemoresistance: enhanced anti-tumor therapy through increased apoptosis by inhibition of NF-kappaB. *Nature Med.* **5**, 412–417 (1999).

14. Jamieson, B. D. & Zack, J. A. In vivo pathogenesis of human immunodeficiency Virus Type 1 reporter virus. *J. Virol.* **72**, 6520–6526 (1998).
15. Chambers, C. A. The expanding world of co-stimulation: the two-signal model revisited. *Trends Immunol.* **22**, 217–223 (2001).
16. Frauwirth, K. A. & Thompson, C. B. Activation and inhibition of lymphocytes by costimulation. *J. Clin. Invest.* **109**, 295–299 (2002).
17. Gruber, M. E., Weih, K. A., Boone, E. J., Smith, P. D. & Clouse, K. A. Endogenous macrophage CSF production is associated with viral replication in HIV-1-infected human monocyte-derived macrophages. *J. Immunol.* **154**, 5528–5535 (1995).
18. Eckstein, D. A. *et al.* HIV-1 Vpr enhances viral burden by facilitating infection of tissue macrophages but not nondividing CD4⁺ T cells. *J. Exp. Med.* **194**, 1407–1419 (2001).
19. Lisignoli, G. *et al.* In vitro cultured stromal cells from human tonsils display a distinct phenotype and induce B cell adhesion and proliferation. *Eur. J. Immunol.* **26**, 17–27 (1996).
20. Stevenson, M., Stanwick, T. L., Dempsey, M. P. & Lamonica, C. A. HIV-1 replication is controlled at the level of T cell activation and proviral integration. *EMBO J.* **9**, 1551–1560 (1990).
21. Zack, J. A. *et al.* HIV-1 entry into quiescent primary lymphocytes: molecular analysis reveals a labile, latent viral structure. *Cell* **61**, 213–222 (1990).
22. Pierson, T. C. *et al.* Molecular characterization of preintegration latency in human immunodeficiency virus type 1 infection. *J. Virol.* **76**, 8518–8531 (2002).
23. Wu, Y. & Marsh, J. W. Selective transcription and modulation of resting T cell activity by preintegrated HIV DNA. *Science* **293**, 1503–1506 (2001).
24. Geyer, M., Fackler, O. T. & Peterlin, B. M. Structure–function relationships in HIV-1 Nef. *EMBO Rep.* **2**, 580–585 (2001).
25. Greene, W. C. & Peterlin, B. M. Charting HIV's remarkable voyage through the cell: Basic science as a passport to future therapy. *Nature Med.* **8**, 673–680 (2002).
26. Davis, S. J. & van der Merwe, P. A. CD2: an exception to the immunoglobulin superfamily concept? *Science* **273**, 1241–1242 (1996).
27. Zhang, Z. *et al.* Sexual transmission and propagation of SIV and HIV in resting and activated CD4⁺ T cells. *Science* **286**, 1353–1357 (1999).
28. Kalter, D. C. *et al.* Enhanced HIV replication in macrophage colony-stimulating factor-treated monocytes. *J. Immunol.* **146**, 298–306 (1991).

Acknowledgements We thank A. Dauphin, and K. Triques for research support, B. Blais for FACS analysis, B. Mellor for preparation of the figures, and T. Pinkos and N. Nelson for manuscript preparation. We also wish to acknowledge assay support provided by the University of Massachusetts Center for AIDS Research. The recombinant IκBα expression plasmid was kindly provided by R. Gaynor; the IκBαSR and LacZ adenovirus vectors were kindly provided by A. Baldwin. HIV-1_{SFII} and HIV-1_{HSA} were obtained through the AIDS Research and Reference Reagent Program, Division of AIDS, NIAID, NIH. This study was supported by grants from the NIH and the Jenner Foundation to M.S.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and request for materials should be addressed to M.S. (Mario.Stevenson@umassmed.edu).

JNK phosphorylates paxillin and regulates cell migration

Cai Huang^{*†}, Zenon Rajfur^{*†}, Christoph Borchers[‡], Michael D. Schaller^{*†§} & Ken Jacobson^{*†§}

^{*} Department of Cell and Developmental Biology, [†] Comprehensive Center for Inflammatory Disorders, [‡] Department of Biochemistry and Biophysics, and [§] Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, North Carolina 27599-7090, USA

The c-Jun amino-terminal kinase (JNK) is generally thought to be involved in inflammation, proliferation and apoptosis^{1,2}. Accordingly, its substrates are transcription factors and anti-apoptotic proteins². However, JNK has also been shown to be required for *Drosophila* dorsal closure^{3,4}, and MAP kinase/ERK kinase 1, an upstream kinase in the JNK pathway, has been shown to be essential for cell migration^{5,6}. Both results imply that JNK is important in cell migration. Here we show that JNK1 is required for the rapid movement of both fish keratocytes and rat bladder tumour epithelial cells (NBT-II). Moreover, JNK1 phosphorylates serine 178 on paxillin, a focal adhesion adaptor, both *in vitro* and in intact cells. NBT-II cells expressing the Ser 178 → Ala mutant of paxillin (Pax_{S178A}) formed focal adhesions and exhibited the limited movement associated with such

contacts in both single-cell-migration and wound-healing assays. In contrast, cells expressing wild-type paxillin moved rapidly and retained close contacts as the predominant adhesion. Expression of Pax_{S178A} also inhibited the migration of two other cell lines. Thus, phosphorylation of paxillin by JNK seems essential for maintaining the labile adhesions required for rapid cell migration.

To explore the role of JNK in cell movement, we employed a time-lapse single-cell-motility assay to test the effect of SP600125, a specific inhibitor of JNK, on the migration of fast-moving fish keratocytes. As shown in Fig. 1a, b, SP600125 inhibited the rapid, directed movement of keratocytes, whereas SB20350 and PD98059, inhibitors of p38 mitogen-activated protein kinase (MAPK) and p42/44MAPK pathways, respectively, had little effect on the migration. This indicates that JNK might be important for the migration of keratocytes.

To confirm the role of JNK in cell migration, we investigated the effects of dominant-negative mutants on cell movement. However, it is currently difficult to transfect keratocytes. We found that NBT-II cells were easily transfected and exhibited rapid keratocyte-like movement when grown in DMEM/F-12 medium (Fig. 1c) and SP600125 also arrested the movement of NBT-II cells (Fig. 1d). Overexpression of JNK1AF, a dominant-negative mutant of JNK1, strongly inhibited the movement of NBT-II cells (Fig. 1e, bottom right), whereas neither p38αAF nor p38βAF, dominant-negative mutants of p38MAPK-α and p38MAPK-β, respectively, had significant effects on the movement of NBT-II cells (Fig. 1e, top right and bottom left). SP600125 also inhibited the migration of Swiss 3T3 fibroblasts into wounds (Supplementary Information 1). Thus, JNK is essential for robust migration in these cell types.

The substrates identified for JNK are transcription factors and anti-apoptotic proteins². Although JNK-mediated changes in transcription are likely to affect migration⁷, transcriptional activity should not be important in the JNK inhibitor experiments because of the short durations involved, suggesting that an unknown substrate for JNK might have a function in JNK-mediated cell movement. Activated (phosphorylated) JNK localizes in focal adhesions⁸. Because paxillin is an adaptor protein involved in adhesion organization and cell migration^{9–11} and it is phosphorylated at serine residues in cells in response to fibronectin and growth factors^{12,13}, we proposed that paxillin is a substrate for JNK.

To determine whether this was so, the phosphorylation of recombinant paxillin-β by JNK1 was evaluated *in vitro*. As shown in Fig. 2a, MAPK kinase (MKK)7-activated JNK1 strongly phosphorylated paxillin. Non-activated JNK1 caused a slight phosphorylation of paxillin, and MKK7 alone did not induce phosphorylation. This result indicates that paxillin is a substrate for JNK *in vitro*. Phospho-amino-acid analysis demonstrated that paxillin is phosphorylated on serine residues by JNK1 (Fig. 2b) and phosphopeptide mapping analysis indicated that the major phosphorylation site is located between residues 155 and 391 (Fig. 2c).

To ascertain the major residue targeted by JNK, the phosphopeptide at the major spot was recovered from the cellulose plate and subjected to mass spectrometry. The mass spectrum showed that the parent ion was observed at *m/z* 1945.8733, corresponding to peptide ANSSPPLPGALSPLYGVPE with one phospho group. Two fragment ions were also detected at *m/z* 1279.6237 and 1586.8096, which are equivalent to the *m/z* of peptides [PGALSPLYGVPE]_p and [PPLPGALSPLYGVPE]_p, respectively, indicating that the phosphorylation site is located on peptide PGALSPLYGVPE. Because the only serine residue in the peptide PGALSPLYGVPE is Ser 178, this is the major phosphorylation site. This conclusion was confirmed by site-directed mutagenesis. As shown in Fig. 2d, e, the replacement of Ser 178 with Ala to form the mutant Pax_{S178A} markedly inhibited paxillin phosphorylation, and the major spot is absent from the phosphopeptide map of Pax_{S178A}. Thus, Ser 178 of paxillin is the major residue targeted by JNK *in vitro*.

We next considered whether paxillin is also a substrate for JNK in intact cells. HEK-293 cells were transfected with paxillin- β tagged with enhanced green fluorescent protein (EGFP), with or without JNK1 co-transfection, and labelled with [32 P]orthophosphoric acid. EGFP-paxillin- β was immunoprecipitated and subjected to phosphorylation analysis. As shown in Fig. 3a, phosphorylation of paxillin- β was enhanced by co-transfection with JNK1 (lanes 2 and 3). The paxillin bands from Fig. 3a were analysed by phosphopeptide mapping. Two major spots (a and b) were observed on the maps of EGFP-paxillin- β , and co-transfection of JNK1 caused an increase in the intensity of the spots (Fig. 3b, left two panels). Spot b also migrated with the major spot from the map of paxillin- β phosphorylated *in vitro* (Fig. 3b, right two panels). In addition, paxillin phosphorylation was stimulated by epidermal growth factor (EGF), a physiological stimulator of JNK and cell migration, and inhibited by the JNK inhibitor SP600125 (see Fig. 3d). Paxillin is therefore also a substrate for JNK in cells, and the same peptide is phosphorylated by JNK *in vitro* and in cells.

To clarify whether Ser 178 is also the major phosphorylation site in cells, NBT-II cells expressing EGFP-paxillin- β or EGFP-Pax_{S178A} were transfected with JNK1. The cells were labelled and the phosphorylation of EGFP-paxillin- β and EGFP-Pax_{S178A} was determined as above. As expected, transfection of JNK1 stimulated the phosphorylation of EGFP-paxillin- β but not that of EGFP-Pax_{S178A} (Fig. 3c). The mutation at Ser 178 had no

effect on the tyrosine phosphorylation of paxillin (Supplementary Information 2). This result indicates that Ser 178 is also the major residue targeted by JNK in cells. Previously, Ser 188 and Ser 190 were identified as main phosphorylation sites on avian paxillin¹², but these sites are not conserved on mammalian paxillin. LIM domain 2 conjugated with glutathione S-transferase was also phosphorylated *in vitro* at Thr 403 by a LIM-domain-associated kinase, and mutation at this residue was shown to impair cell adhesion¹⁴. In contrast, Ser 178 is conserved in different mammalian paxillin isoforms.

To determine whether endogenous paxillin is also phosphorylated at Ser 178 by physiological stimuli, MDA-MB-231 human breast-cancer cells were labelled and treated with EGF, a physiological stimulator of JNK (Supplementary Information 3), and the phosphorylation of Ser 178 on endogenous paxillin was determined. As shown in Fig. 3d, the phosphorylation of paxillin was stimulated by EGF with a similar time course as that of JNK activation, and the stimulation was inhibited by the JNK inhibitor SP600125. Further phosphopeptide mapping analysis indicated that the phosphorylation of Ser 178 was significantly upregulated by EGF and downregulated by SP600125 (Fig. 3e). The phosphorylation of paxillin at Ser 178 might therefore mediate certain physiological functions in response to stimulation by EGF.

To assess the role of JNK-mediated paxillin phosphorylation in cell migration, NBT-II cells were transfected with EGFP-paxillin- β or EGFP-Pax_{S178A}, and cell migration was measured with the time-lapse single-cell-motility assay. As shown in Fig. 4a, whereas most of the cells expressing EGFP-Pax_{S178A} lose polarity and exhibit markedly reduced migration (right), cells expressing wild-type EGFP-paxillin- β , like parental NBT-II cells, have a keratocyte-like morphology and exhibit extensive migration (left). Similar results

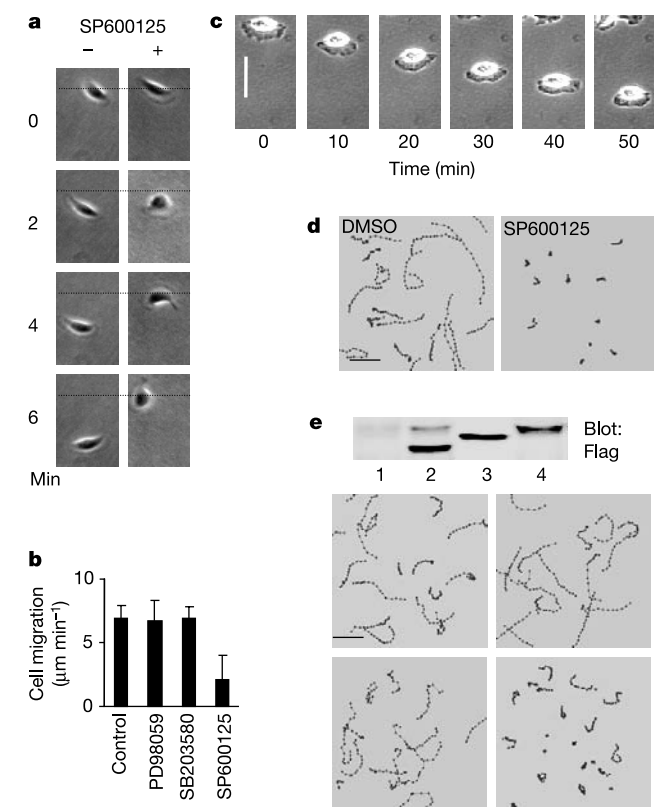


Figure 1 JNK1 is required for cell migration. **a**, The movement of fish keratocytes before and after the addition of 50 μ M SP600125 (Calbiochem). **b**, Inhibition of keratocyte movement by SP600125, but not by PD98059 and SB203580, at 25 μ M. **c**, The keratocyte-like movement of a typical NBT-II cell. Scale bar, 50 μ m. **d**, Inhibition of the movement of NBT-II cells by SP600125 (50 μ M). Scale bar, 110 μ m. **e**, Inhibition of the movement of NBT-II cells by stably transfecting Flag-JNK1AF, but not by Flag-p38 α AF or Flag-p38 β AF (lower panels). The expression of individual clones was probed with anti-flag antibody (upper panels). Lane 1 and top left panel, non-transfected; lane 2 and top right panel, p38 α AF; lane 3 and bottom left panel, p38 β AF; lane 4 and bottom right panel, JNK1AF. Scale bar, 110 μ m.

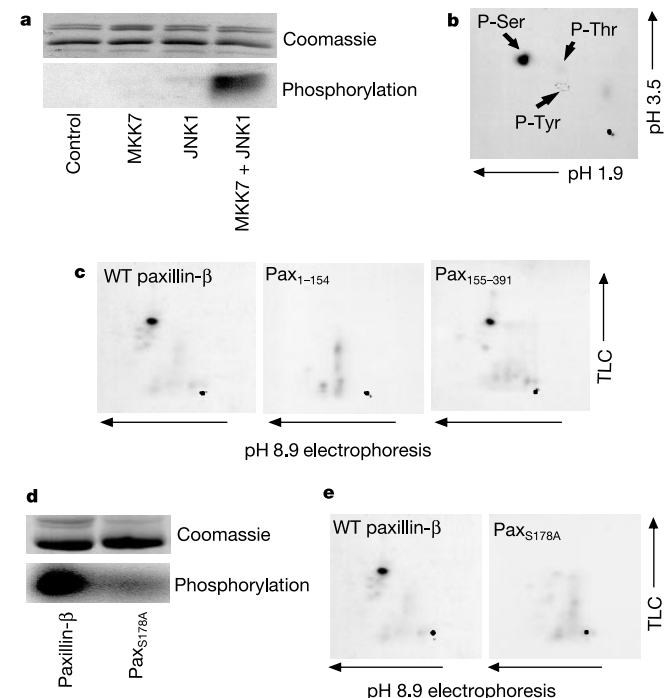


Figure 2 Paxillin is phosphorylated at Ser 178 *in vitro*. **a**, Direct phosphorylation of paxillin by JNK1 *in vitro*. His-tagged paxillin- β was incubated respectively with active MKK7 (Upstate), JNK1 and MKK7 + JNK1 in 50 μ l kinase buffer containing 40 μ M [γ - 32 P]ATP (10 μ Ci). **b**, Phospho-amino-acid analysis shows predominantly phosphoserine. **c**, Phosphopeptide mapping analysis of wild-type (WT) paxillin- β , Pax₁₋₁₅₄, and Pax₁₅₅₋₃₉₁. **d**, His-tagged wild-type paxillin- β and Pax_{S178A} were phosphorylated with preactivated JNK1 (0.2 μ g). **e**, Phosphopeptide mapping analysis of wild-type paxillin- β and Pax_{S178A}. The origins are marked by dots.

were obtained with a wound-healing assay to examine cell motility (Supplementary Information 4). Thus, phosphorylation of paxillin at Ser 178 is required for the migration of single NBT-II cells and also for the directed migration of cells to close a wound. Because the expression of EGFP-Pax_{S178A} also inhibited the directed migration of the MDA-MB-231 human breast cancer cells (Fig. 4b) and Chinese hamster ovary K1 (Fig. 4c) cells into wounds, phosphorylation of paxillin at Ser 178 might be a key event in the regulation of migration in different cell types.

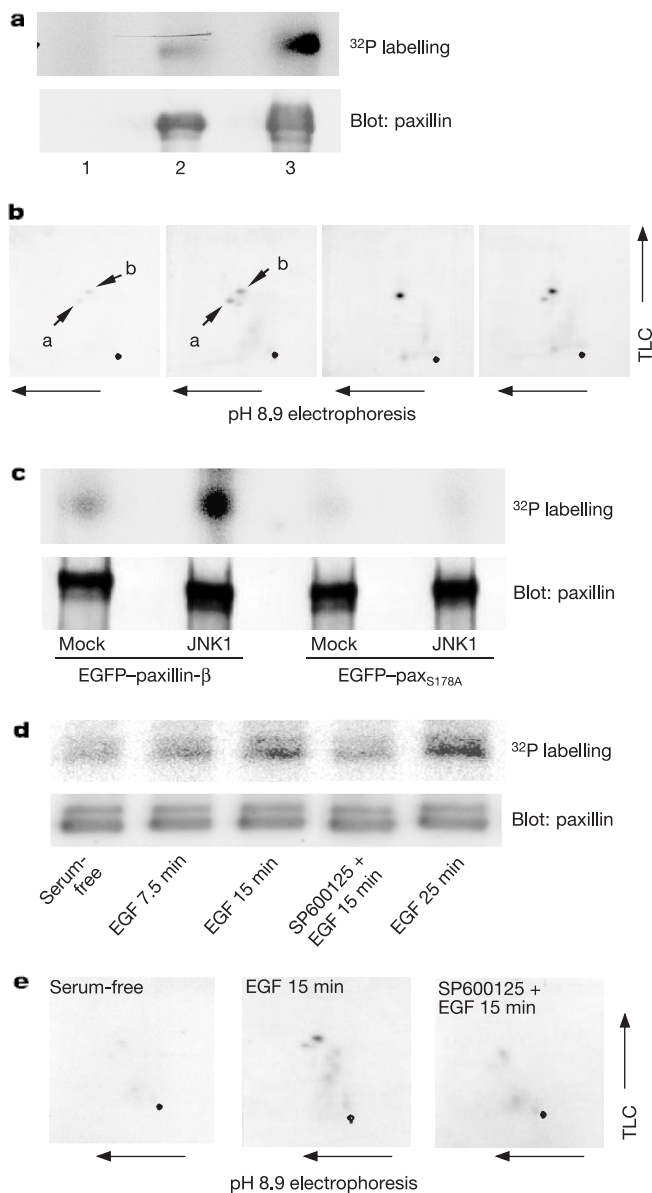


Figure 3 Paxillin is also phosphorylated at Ser 178 in cells. **a**, Overexpression of JNK1 promotes the phosphorylation of paxillin-β in HEK-293 cells. Lane 1, mock; lane 2, EGFP-paxillin-β; lane 3, EGFP-paxillin-β plus JNK1. **b**, Phosphopeptide mapping analysis of paxillin-β phosphorylated in cells and *in vitro*. Left two panels, EGFP-paxillin-β from lanes 2 and 3 of **a**, respectively; third panel, His-tagged paxillin phosphorylated by JNK1 *in vitro*; fourth panel, a mixture of the middle two panels. TLC, thin-layer chromatography. **c**, Abolition of paxillin phosphorylation by replacing Ser 178 with Ala. Cells expressing EGFP-paxillin-β or -Pax_{S178A} were transfected with JNK1 as indicated. **d**, Phosphorylation of endogenous paxillin was stimulated by EGF. MDA-MB-231 cells were labelled, treated with SP600125 (50 μM) and stimulated with EGF (20 ng ml⁻¹) as indicated. **e**, Phosphopeptide mapping analysis of samples from **d**.

Because paxillin is a focal adhesion adaptor protein, it is possible that serine phosphorylation of paxillin might control cell movement through the regulation of adhesion dynamics. This possibility prompted us to examine the adhesions of cells expressing EGFP-paxillin or EGFP-Pax_{S178A}. By using total internal reflection fluorescence (TIRF) microscopy to reveal adhesive structures near the substrate, we found that EGFP-Pax_{S178A} (Fig. 5a, right) was localized in more discrete, focal-adhesion-like structures (bottom right), whereas wild-type EGFP-paxillin-β (Fig. 5a, left) retained a

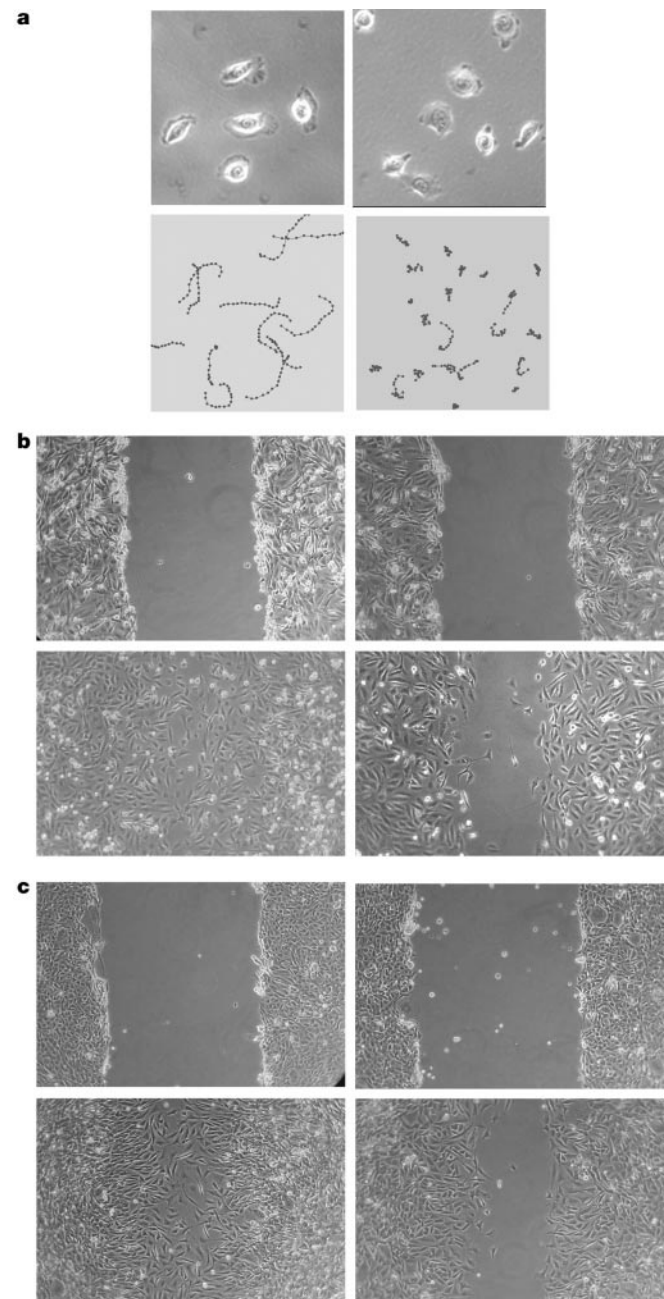


Figure 4 Expression of Pax_{S178A} inhibits cell migration. **a**, Expression of Pax_{S178A} changes cell morphology (top panels) and inhibits cell movement (bottom panels). NBT-II cells were stably expressing EGFP-paxillin-β (left panels) or EGFP-Pax_{S178A} (right panels). **b**, **c**, Inhibition of the migration of MDA-MB-231 (**b**) and Chinese hamster ovary K1 (**c**) cells by expressing Pax_{S178A}. Confluent monolayers of cells stably expressing EGFP-paxillin-β (left panels) or EGFP-Pax_{S178A} (right panels) were wounded and images were taken when the wounds were made (top panels) and after incubation for 15 h (bottom panels).

more diffuse distribution with few punctate structures at the cell edges (bottom left). In addition, phalloidin and immunofluorescence staining showed (Fig. 5b and Supplementary Information 5) that cells expressing EGFP-Pax_{S178A} (Fig. 5b, right) formed peripheral stress fibres (bottom right) and accumulated vinculin (top right) at points along the stress fibers. By contrast, cells expressing EGFP-paxillin- β (Fig. 5b, left) exhibited randomly distributed tiny vinculin dots (top left) and had few stress fibres (bottom left). In addition, we found that overexpression of the JNK1 dominant-negative mutant induced peripheral vinculin accumulation and focal adhesion formation (C.H. and K.J., unpublished observations). However, the difference in focal adhesions between NBT-II cells transfected with EGFP-paxillin (Fig. 5a, bottom left) and with EGFP-Pax_{S178A} (Fig. 5a, bottom right) is not observed in less motile Swiss 3T3 cells expressing EGFP-paxillin or EGFP-Pax_{S178A} (Supplementary Information 6). Because large focal adhesions normally do not undergo rapid turnover, tending to be involved passively in anchorage¹⁵, and fast-moving cells usually do not form focal adhesions¹⁶, focal adhesion formation might be responsible for the inhibition of cell movement when Pax_{S178A} is expressed.

Activation of JNK is correlated with an increase in cell migration in several systems^{17–19}. Consistent with this is the observation that JNK is a common downstream target of multiple signalling pathways controlling cell movement. Rac, a small GTP-binding protein

essential for lamellipodia formation and cell movement²⁰, has been shown to activate JNK²¹. MAP kinase/ERK kinase 1, an upstream kinase of JNK, is associated with both α -actinin²² and Rac²³ and is crucial to cell movement^{5,6}. Focal adhesion kinase, which is required for cell movement²⁴, has been shown to transmit integrin-stimulated signals to JNK^{8,25}. In addition, Src, a regulator of cell migration²⁶, is essential for signal transduction from integrins to JNK²⁵. Because JNK-mediated phosphorylation of paxillin on Ser 178 regulates cell adhesion and movement, it is possible that these pathways all exert their influence on cell adhesion and movement by means of the JNK-paxillin connection. Moreover, because JNK is associated with microtubules²⁷ that in turn affect focal adhesion turnover²⁸, the possible involvement of microtubules in the JNK-paxillin regulation of cell migration should not be overlooked. $\square$

Methods

Cell migration

For keratocyte movement, fish keratocytes were isolated as described previously²⁹. The cells were plated on Petri dishes coated with 20 $\mu\text{g ml}^{-1}$ fibronectin and incubated at 28 °C for 1 h. Cell motility was recorded as described previously²⁹. For the movement of NBT-II cells, bacterial Petri dishes (35 mm) were coated by incubating 20 $\mu\text{g ml}^{-1}$ collagen for 1 h. NBT-II cells (A.T.C.C.) were treated with trypsin and resuspended in DMEM/F12 medium containing 10% fetal bovine serum, plated at low density on the dishes, and cultured for 12 h in a CO₂ incubator. Cell motility was measured with a Zeiss microscope equipped with a Hamamatsu C4880 charge-coupled device (CCD) camera and an open perfusion micro-incubator (Harvard Apparatus). Migration was tracked for 3 h; images were recorded every 10 min. The movement of individual cells was analysed with Metamorph software (Universal Imaging). Wound healing assays were performed as described previously³⁰. Mitomycin C (1 $\mu\text{g ml}^{-1}$) was included to inhibit cell proliferation.

Plasmids and transfection

Flag-p38 α AF and Flag-p38 β AF in pcDNA3.1 vector were gifts from J. Han (Scripps Research Institute). Flag-JNK1 was obtained from K. Yoshioka (Kanazawa University, Japan). Human paxillin cDNA was from H. Sabe (Osaka Bioscience Institute, Japan). JNK1AF (S183A and Y185F) and Pax_{S178A} were created by *Pfu*-based polymerase chain reaction with a Quick-Change mutation kit (Stratagene). His-tagged paxillin- β , Pax_{155–391} and JNK1 were generated by subcloning PCR-amplified DNA fragments into pQE-30 vector (Qiagen). His-tagged Pax_{1–154} was created by sequentially digesting pQE-paxillin- β with *Bsr*G1, *Bam*H1 and mung bean nuclease, and circularizing the larger fragment. EGFP-paxillin- β and EGFP-Pax_{S178A} were generated by subcloning DNA fragments expressing wild-type paxillin and Pax_{S178A} into pEGFP-C vector (Clontech). His-tagged proteins were purified by affinity chromatography with Ni²⁺-nitrilotriacetate-agarose. Cells were transfected with Lipofectamine Plus transfection reagent (Invitrogen). Stable transfected cells were obtained by screening the G418-resistant clones. At least three independent set of clones were examined in experiments.

Phosphorylation *in vitro*

His-tagged paxillin- β (3 μg) was incubated with kinase ($\sim 0.5 \mu\text{g}$) in 50 μl of kinase buffer (20 mM MOPS pH 7.2, 10 mM MgCl₂, 25 mM β -glycerol phosphate, 5 mM EGTA, 1 mM sodium orthovanadate, 1 mM dithiothreitol) containing 40 μM ATP (10 μCi [γ -³²P]ATP) for 30 min at 24 °C. The reactions were terminated by adding SDS-sample buffer. The samples were subjected to SDS-polyacrylamide-gel electrophoresis and transferred to a nitrocellulose membrane (polyvinylidene difluoride (PVDF) membrane when phospho-amino-acids were analysed) for autoradiography.

Phosphorylation in intact cells

Cells were incubated with [³²P]orthophosphoric acid in sodium-phosphate-deficient DMEM. After 12 h, cells were harvested and lysed with a modified RIPA buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 1% IPEGAL, 0.5% deoxycholate, 5 mM EDTA, 1 mM phenylmethylsulphonyl fluoride, 10 $\mu\text{g ml}^{-1}$ aprotinin, 10 $\mu\text{g ml}^{-1}$ leupeptin, 10 $\mu\text{g ml}^{-1}$ pepstatin, 500 nM okadaic acid, 5 mM α -naphthyl acid phosphate). The lysates were immunoprecipitated with anti-GFP polyclonal antibody (Clontech) or anti-paxillin monoclonal antibody (Upstate Biotechnology). The immune complexes were analysed by SDS-polyacrylamide-gel electrophoresis and transferred to nitrocellulose membrane for the detection of phosphorylation by autoradiography.

Phospho-amino-acid analysis and peptide mapping

Phospho-amino-acid analysis was performed as described previously³⁰. For peptide mapping, protein bands cut from nitrocellulose membrane were digested for 8 h with endoproteinase Glu-C (Sigma) in ammonium bicarbonate pH 7.8 at 37 °C. The samples were dried in a Speed-Vac, then spotted onto a cellulose plate for two-dimensional phosphopeptide mapping, with pH 8.9 buffer for electrophoresis and phosphor-chromatography buffer for thin-layer chromatography.

EGFP fluorescence, TIRF and immunofluorescence

Cells expressing EGFP-paxillin- β or EGFP-Pax_{S178A} were plated on MatTek glass-bottomed dishes precoated with collagen, and cultured for 12 h. EGFP fluorescence and

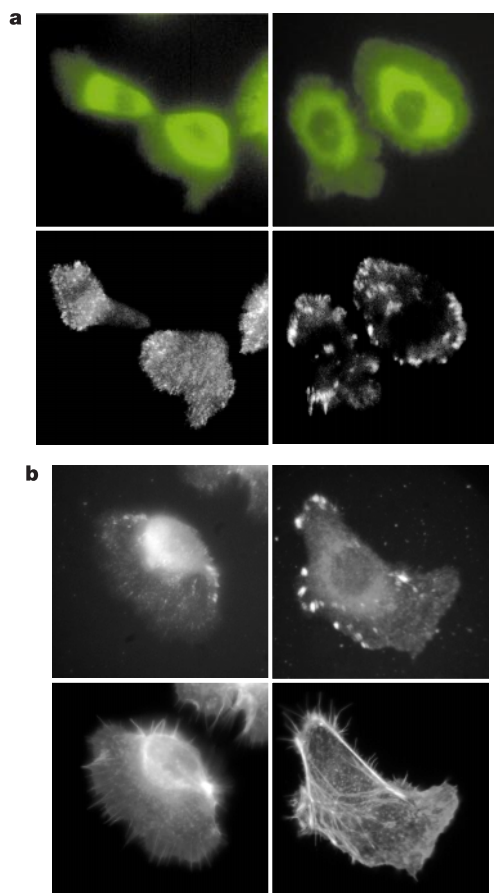


Figure 5 Expression of Pax_{S178A} promotes focal adhesion formation. **a**, Epifluorescence (top panels) and TIRF (bottom panels) images were taken from NBT-II cells stably expressing EGFP-paxillin- β (left panels) and EGFP-Pax_{S178A} (right panels). **b**, Accumulation of peripheral stress fibres and vinculin in cells expressing Pax_{S178A}. NBT-II cells stably expressing EGFP-paxillin- β (left panels) and EGFP-Pax_{S178A} (right panels) were fixed and then stained for vinculin (top panels) and F-actin (bottom panels).

TIRF images were taken with an Olympus IX81 microscope equipped with a 60 $\times$, 1.45 numerical aperture objective and a Hamamatsu C4880 CCD camera. For immunofluorescence staining, NBT-II cells stably expressing EGFP-paxillin- β or EGFP-Pax_{5178A} were plated on collagen-coated plastic rectangles (25 mm $\times$ 50 mm, cut from the bottom of Petri dishes) and cultured for 12 h. F-actin was stained with Alexa 488 phalloidin, and vinculin was stained with anti-vinculin monoclonal antibody (Sigma), followed by tetramethylrhodamine β -isothiocyanate-labelled anti-mouse antibody. The plastic rectangles were then covered with coverslips. F-actin and vinculin images were observed as described above.

Received 15 April 2002; accepted 13 May 2003; doi:10.1038/nature01745.

1. Ip, Y. T. & Davis, R. J. Signal transduction by the c-Jun N-terminal kinase (JNK)—from inflammation to development. *Curr. Opin. Cell Biol.* **10**, 205–219 (1998).
2. Barr, R. K. & Bogoyevitch, M. A. The c-Jun N-terminal protein kinase family of mitogen-activated protein kinases (JNK MAPKs). *Int. J. Biochem. Cell Biol.* **33**, 1047–1063 (2001).
3. Riesgo-Escovar, J. R., Jenni, M., Fritz, A. & Hafen, E. The *Drosophila* Jun-N-terminal kinase is required for cell morphogenesis but not for Djun-dependent cell fate specification in the eye. *Genes Dev.* **10**, 2759–2768 (1996).
4. Sluss, H. K., Han, Z., Barrett, T., Davis, R. J. & Ip, Y. T. A JNK signal transduction pathway that mediates morphogenesis and an immune response in *Drosophila*. *Genes Dev.* **10**, 2745–2758 (1996).
5. Xia, Y. *et al.* MEK kinase 1 is critically required for c-Jun N-terminal kinase activation by proinflammatory stimuli and growth factor-induced cell migration. *Proc. Natl Acad. Sci. USA* **97**, 5243–5248 (2000).
6. Yujiri, T. *et al.* MEK kinase 1 gene disruption alters cell migration and c-Jun NH₂-terminal kinase regulation but does not cause a measurable defect in NF- κ B activation. *Proc. Natl Acad. Sci. USA* **97**, 7272–7277 (2000).
7. Ozanne, B. W. *et al.* Transcriptional regulation of cell invasion: AP-1 regulation of a multigenic invasion programme. *Eur. J. Cancer* **36**, 1640–1648 (2000).
8. Almeida, E. A. *et al.* Matrix survival signaling: From fibronectin via focal adhesion kinase to c-Jun NH₂-terminal kinase. *J. Cell Biol.* **149**, 741–754 (2000).
9. Liu, S. *et al.* Binding of paxillin to α 4 integrins modifies integrin-dependent biological responses. *Nature* **402**, 676–681 (1999).
10. Hagel, M. *et al.* The adaptor protein paxillin is essential for normal development in the mouse and is a critical transducer of fibronectin signaling. *Mol. Cell Biol.* **22**, 901–915 (2002).
11. Schaller, M. D. Paxillin: A focal adhesion-associated adaptor protein. *Oncogene* **20**, 6459–6472 (2001).
12. Bellis, S. L., Perrotta, J. A., Curtis, M. S. & Turner, C. E. Adhesion of fibroblasts to fibronectin stimulates both serine and tyrosine phosphorylation of paxillin. *Biochem. J.* **325**, 375–381 (1997).
13. Vadlamudi, R., Adam, L., Talukder, A., Mendelsohn, J. & Kumar, R. Serine phosphorylation of paxillin by heregulin- β 1: role of p38 mitogen activated protein kinase. *Oncogene* **18**, 7253–7264 (1999).
14. Brown, M. C., Perrotta, J. A. & Turner, C. E. Serine and threonine phosphorylation of the paxillin LIM domains regulates paxillin focal adhesion localization and cell adhesion to fibronectin. *Mol. Biol. Cell.* **9**, 1803–1816 (1998).
15. Beningo, K. A., Dembo, M., Kaverina, I., Small, J. V. & Wang, Y. L. Nascent focal adhesions are responsible for the generation of strong propulsive forces in migrating fibroblasts. *J. Cell Biol.* **153**, 881–888 (2001).
16. Lee, J. & Jacobson, K. The composition and dynamics of cell–substratum adhesions in locomoting fish keratocytes. *J. Cell Sci.* **110**, 2833–2844 (1997).
17. Abassi, Y. A. & Vuori, K. Tyrosine 221 in Crk regulates adhesion-dependent membrane localization of Crk and Rac and activation of Rac signaling. *EMBO J.* **21**, 4571–4582 (2002).
18. Hauck, C. R. *et al.* Inhibition of focal adhesion kinase expression or activity disrupts epidermal growth factor-stimulated signaling promoting the migration of invasive human carcinoma cells. *Cancer Res.* **61**, 7079–7090 (2001).
19. Huynh-Do, U. *et al.* Ephrin-B1 transduces signals to activate integrin-mediated migration, attachment and angiogenesis. *J. Cell Sci.* **115**, 3073–3081 (2002).
20. Ridley, A. J., Allen, W. E., Peppelenbosch, M. & Jones, G. E. Rho family proteins and cell migration. *Biochem. Soc. Symp.* **65**, 111–123 (1999).
21. Minden, A., Lin, A., Claret, F. X., Abo, A. & Karin, M. Selective activation of the JNK signaling cascade and c-Jun transcriptional activity by the small GTPases Rac and Cdc42Hs. *Cell* **81**, 1147–1157 (1995).
22. Christerson, L. B., Vanderbilt, C. A. & Cobb, M. H. MEK1 interacts with alpha-actinin and localizes to stress fibers and focal adhesions. *Cell Motil. Cytoskel.* **43**, 186–198 (1999).
23. Fanger, G. R., Johnson, N. L. & Johnson, G. L. MEK kinases are regulated by EGF and selectively interact with Rac/Cdc42. *EMBO J.* **16**, 4961–4972 (1997).
24. Ilic, D. *et al.* Reduced cell motility and enhanced focal adhesion contact formation in cells from FAK-deficient mice. *Nature* **377**, 539–544 (1995).
25. Oktay, M., Wary, K. K., Dans, M., Birge, R. B. & Giancotti, F. G. Integrin-mediated activation of focal adhesion kinase is required for signaling to Jun NH₂-terminal kinase and progression through the G1 phase of the cell cycle. *J. Cell Biol.* **145**, 1461–1469 (1999).
26. Liu, J., Huang, C. & Zhan, X. Src is required for cell migration and shape changes induced by fibroblast growth factor 1. *Oncogene* **18**, 6700–6706 (1999).
27. Nagata, K. *et al.* The MAP kinase kinase kinase MLK2 co-localizes with activated JNK along microtubules and associates with kinesin superfamily motor KIF3. *EMBO J.* **17**, 149–158 (1998).
28. Small, J. V. & Kaverina, I. Microtubules meet substrate adhesions to arrange cell polarity. *Curr. Opin. Cell Biol.* **15**, 40–47 (2003).
29. Lee, J., Ishihara, A., Oxford, G., Johnson, B. & Jacobson, K. Regulation of cell movement is mediated by stretch-activated calcium channels. *Nature* **400**, 382–386 (1999).
30. Huang, C., Liu, J., Haudenschild, C. C. & Zhan, X. The role of tyrosine phosphorylation of cortactin in the locomotion of endothelial cells. *J. Biol. Chem.* **273**, 25770–25776 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank C. Raska, J. Han and R. M. Pope for their mass spectrometric analysis, and K. Burridge, L. Graves and C. Otey for critically reading the manuscript. This study was supported by NIH grants to K.J. and M.D.S., the Cell Migration Consortium to K.J. and the National Institute for Dental and Cranial Research grant to K.J. and M.D.S.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to K.J. (frap@med.unc.edu)

Acute mutation of retinoblastoma gene function is sufficient for cell cycle re-entry

Julien Sage*, Abigail L. Miller*, Pedro A. Pérez-Mancera‡, Julianne M. Wysocki* & Tyler Jacks*†

* Department of Biology and Centre for Cancer Research, and † Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

‡ Instituto de Biología Molecular y Celular del Cáncer, CSIC/Universidad de Salamanca, 37007-Salamanca, Spain

Cancer cells arise from normal cells through the acquisition of a series of mutations in oncogenes and tumour suppressor genes¹. Mouse models of human cancer often rely on germline alterations that activate or inactivate genes of interest. One limitation of this approach is that germline mutations might have effects other than somatic mutations, owing to developmental compensation^{2,3}. To model sporadic cancers associated with inactivation of the retinoblastoma (*RB*) tumour suppressor gene in humans, we have produced a conditional allele of the mouse *Rb* gene. We show here that acute loss of *Rb* in primary quiescent cells is sufficient for cell cycle entry and has phenotypic consequences different from germline loss of *Rb* function. This difference is explained in part by functional compensation by the *Rb*-related gene *p107*. We also show that acute loss of *Rb* in senescent cells leads to reversal of the cellular senescence programme. Thus, the use of conditional knockout strategies might refine our understanding of gene function and help to model human cancer more accurately.

To model somatic inactivation of *RB* in human cancer, we have created a conditional allele of the mouse *Rb* gene by using gene targeting in embryonic stem cells (Fig. 1a)⁴. Correct targeting and the presence of *loxP* sites flanking *Rb* exon 3 were verified by polymerase chain reaction (PCR) and Southern blot analysis (data not shown). Conditional *Rb* homozygous mutant mice displayed no developmental defects⁵, and ageing mice did not develop tumours associated with the loss of *Rb* (J.S. and T.J., unpublished observations), indicating that this conditional allele was wild-type for *Rb* function before further deletion by the Cre recombinase. The proliferation rate of conditional *Rb* homozygous mouse embryonic fibroblasts (*cRb*^{lox/lox} MEFs) was indistinguishable from that of wild-type MEFs, and the levels of pRB were identical in *cRb*^{lox/lox} MEFs and in wild-type MEFs (data not shown).

When exponentially growing *cRb*^{lox/lox} MEFs were infected with an adenovirus vector expressing the Cre recombinase (Ad-Cre), *Rb* exon 3 was deleted within 24 h (Fig. 1b). pRB was no longer detectable 3 days after infection with Ad-Cre (Fig. 1c). Levels of the two pRB family members p107 and p130 were not significantly affected by loss of pRB in exponentially growing populations, except

for a slight but reproducible increase in p107 levels (Fig. 1c). Levels of p107 are also elevated in germline *Rb*^{-/-} MEFs⁶. We did not observe any inhibitory or growth-promoting effect of Ad or Ad-Cre infection on the proliferation of wild-type, *Rb*^{-/-} and *cRb*^{lox/lox} cells, indicating that, at the dose used, Ad-Cre is not cytotoxic in this system and that Ad or Ad-Cre does not cooperate with loss of pRB in cell cycle deregulation (Fig. 1d and data not shown). In addition, *Rb*^{-/-} and *cRb*^{lox/lox} MEFs had similar proliferation curves, showing that, in this respect, germline and acute loss of *Rb* function were phenotypically equivalent (Fig. 1e). *Rb*^{-/-} and *cRb*^{lox/lox} MEFs were indistinguishable in several other assays as well, including response to contact inhibition and proliferation at low density (data not shown). In summary, we have developed a system to eliminate *Rb* function acutely in primary MEFs. This system has allowed us to explore conditions under which acute *Rb* mutation might differ from the constitutive absence of *Rb*.

Maintenance of the growth-arrested state (quiescence) is crucial for normal development and differentiation. Failure to maintain quiescence can cause inappropriate proliferation, which can increase the target size for additional cancer-promoting mutations and can directly promote tumorigenesis. We therefore chose to study the consequences of acute loss of *Rb* function in quiescent cells. MEFs were first growth-arrested for 3 days at full confluence without the addition of fresh medium. The confluent cells were then rendered quiescent by leaving them for 3 days in 0.1% serum and for a further 3 days after replating in 0.1% serum at ~50% confluence. Quiescent germline *Rb*^{-/-} MEFs were infected with Ad or Ad-Cre and kept for 3 days in 0.1% serum before incubation for 24 h with 5-bromodeoxyuridine (BrdU). Under these conditions, ~20% of the *Rb*^{-/-} cells incorporated BrdU (Fig. 2a). For quiescent *cRb*^{lox/lox} MEFs, either untreated or Ad-infected, ~10% of the cells incorporated BrdU in a 24-hour period (Fig. 2a). These data indicate that, as reported previously⁷, germline loss of *Rb* function does not prevent cells from arresting in G0, but germline *Rb*^{-/-} cells may be slightly resistant to G0 arrest induced by low serum concentration. Strikingly, however, quiescent *cRb*^{lox/lox} MEFs infected with Ad-Cre

showed a marked increase in the percentage of cells incorporating BrdU (~50% BrdU-positive cells in a 24-hour period; Fig. 2a), showing that acute loss of *Rb* is phenotypically different from its constitutive absence in this assay.

The relative normality of germline *Rb*^{-/-} MEFs might be due to differences in the levels of cell cycle regulators compared with those in quiescent *cRb*^{lox/lox} cells. The candidate cell cycle inhibitors p21^{CIP1} and p16^{INK4a} were expressed at similar levels in quiescent *cRb*^{lox/lox} and *Rb*^{-/-} MEFs (Fig. 2b). In contrast, targets of the E2F transcription factor, whose activity is deregulated in *Rb* mutant cells⁸, were significantly upregulated in quiescent *Rb*^{-/-} MEFs (Fig. 2b) compared with *cRb*^{lox/lox} MEFs. However, some of these E2F target genes, such as *E2f-1* itself and the *cyclin E* gene, would be expected to promote cell cycle entry, and their upregulation cannot explain why germline *Rb* mutant cells were capable of entering and maintaining quiescence. The E2F target genes and cell cycle inhibitors p19^{ARF} and p107 were also markedly upregulated in quiescent *Rb*^{-/-} MEFs compared with *cRb*^{lox/lox} cells (Fig. 2b, c). Northern blot analysis (Fig. 2d) and real-time PCR (not shown) showed that p107 upregulation occurred largely at the transcriptional level, probably as a result of increased E2F activity in these cells. p130 protein levels were similar in the two populations (Fig. 2c).

To test the hypothesis that elevated levels of p107 in quiescent *Rb*^{-/-} cells were at least partly responsible for their growth-arrested state, we sought to downregulate p107 expression by the stable infection of exponentially dividing *Rb*^{-/-} MEF populations with a retroviral vector allowing the production of small interfering RNA (siRNA) molecules directed against p107 ('si-p107'). si-p107 *Rb*^{-/-} MEFs expressed barely detectable levels of p107 but unchanged levels of p130 (Fig. 2e). The role of p107 in controlling the entry of *Rb*^{-/-} MEFs in quiescence was first assayed by comparing proliferation indices in cells 24 h after the removal of growth factors (0.1% serum) with cells grown in 10% serum using incubation with BrdU for 1 h. At low serum concentration, the number of *Rb*^{-/-} MEFs infected with a control retrovirus that incorporated BrdU decreased 2.9 ± 0.5-fold (mean ± s.d., *n* = 7) compared with cells

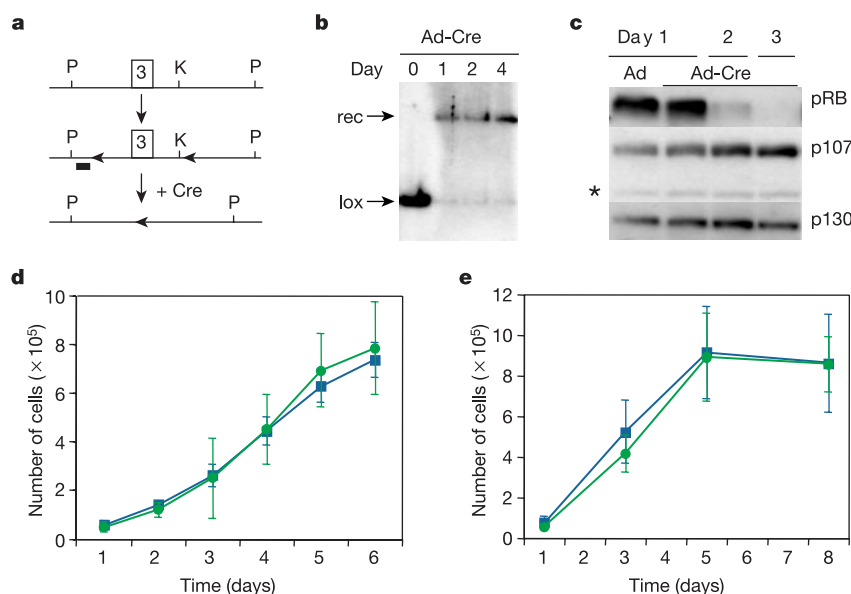


Figure 1 Acute deletion of *Rb* in MEFs. **a**, Creation of a conditional allele of *Rb*. After targeting, *Rb* exon 3 is flanked by loxP sites (arrow heads). After Cre-mediated recombination, *Rb* exon 3 is deleted, leaving one loxP site in the genome. P, PstI, K, KpnI. **b**, Southern blot analysis of *Rb* exon 3 deletion after infection of *cRb*^{lox/lox} MEFs with Ad-Cre. lox, conditional allele; rec, recombined deleted allele. Genomic DNA was digested

with PstI and KpnI. The probe used corresponds to the black box in **a**. **c**, Immunoblot analysis of the three pRB-family members after infection of proliferating *cRb*^{lox/lox} MEFs with Ad or Ad-Cre (asterisk, non-specific protein band). **d**, Proliferation curve of *cRb*^{lox/lox} MEFs infected with Ad (squares) or Ad-Cre (circles) (*n* = 5). **e**, Proliferation curve of *cRb*^{lox/lox} (*cRb*^{-/-}, squares) and *Rb*^{-/-} (circles) MEFs infected with Ad-Cre (*n* = 6).

grown at full serum concentration (~30% BrdU-positive cells for all the different populations studied). In contrast, the number of si-p107 $Rb^{-/-}$ cells incorporating BrdU decreased only 1.6 ± 0.4 -fold (mean $\pm$ s.d., $n = 5$) in 0.1% serum compared with 10% serum, indicating that p107 plays a role in the entry of $Rb^{-/-}$ MEFs into quiescence.

We then attempted to make $Rb^{-/-}$ and si-p107 $Rb^{-/-}$ MEF populations fully quiescent after 6 days in 0.1% serum by following our initial protocol. The vast majority of quiescent si-p107 $Rb^{-/-}$ cells died under these conditions (not shown), a phenomenon similar to that previously observed with Rb -family triple knockout cells⁴. To bypass the death of si-p107 $Rb^{-/-}$ MEFs entering full quiescence, we used a lentiviral vector co-expressing a reporter gene encoding green fluorescent protein (GFP) and the same siRNA molecules directed against p107 (ref. 9). In contrast to MLV-derived retroviruses, lentiviruses can infect non-dividing, quiescent cells. p107 knock-down in already quiescent lentivirus-infected $Rb^{-/-}$ MEFs was confirmed by western blot analysis and immunofluorescence (data not shown). Significantly, loss of p107 in quiescent $Rb^{-/-}$ MEFs led to a threefold increase in BrdU incorporation, showing that p107 expression is indeed necessary for the maintenance of quiescence in $Rb^{-/-}$ MEFs (Fig. 2f). Similar experiments with siRNA molecules directed against p19^{ARF} indicated that p19^{ARF} is

also required for the establishment of full quiescence in $Rb^{-/-}$ MEFs (Supplementary Fig. 1).

The fact that levels of p107 are mostly controlled transcriptionally suggested that $cRb^{-/-}$ cells might compensate rapidly for the loss of Rb . This possibility was also indicated by the fact that cycling early-passage $Rb^{-/-}$ and $cRb^{-/-}$ MEFs had otherwise similar properties (see Fig. 1). In fact, we found that infection of $cRb^{lox/lox}$ MEFs at the time when medium was changed to 0.1% serum (6 days before the time that we consider the cells to be quiescent) rendered them similar to $Rb^{-/-}$ MEFs with regard to p107 levels (data not shown) and BrdU incorporation (Fig. 2g). Thus, we conclude that transcriptional upregulation of p107, along with upregulation of p19^{ARF} in quiescent germline $Rb^{-/-}$ cells can functionally compensate for loss of Rb and allows these cells to respond to anti-proliferative signals. In contrast, in cells in which pRB is actively involved in maintaining the quiescent state, such potential compensatory mechanisms are not deployed, allowing an acute loss of Rb to result in marked cell cycle re-entry.

We next examined whether other forms of cell cycle arrest were dependent on the continuous presence of pRB. It has been proposed that cellular senescence is an anti-cancer programme whose endpoint is permanent cell cycle arrest, preventing ageing or mutant cells from undergoing proliferation^{10–12}. $Rb^{-/-}$ MEFs undergo

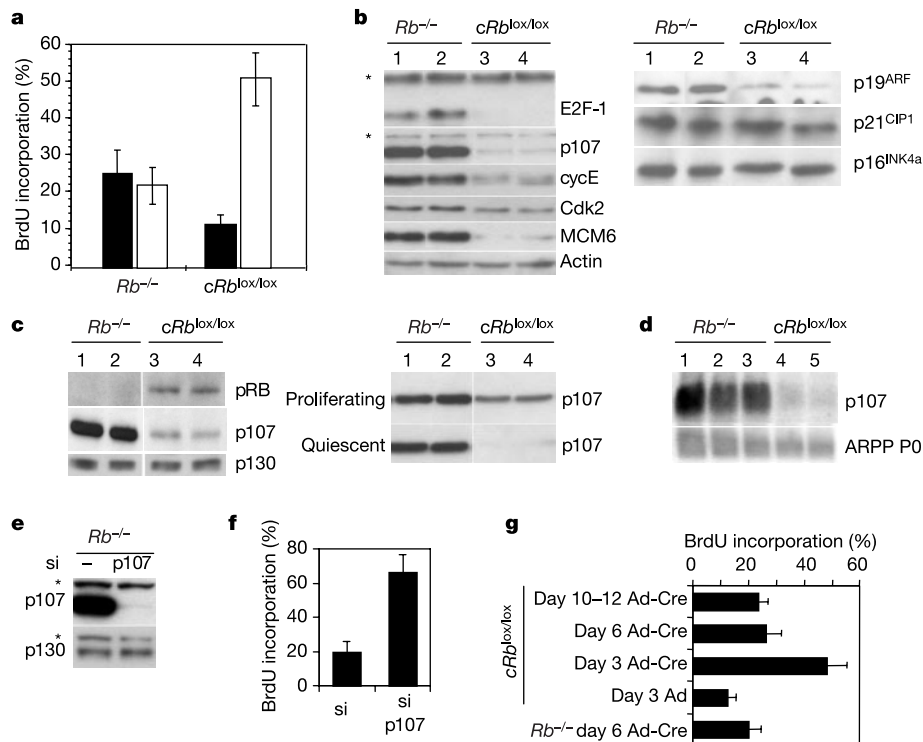


Figure 2 Effects of acute versus germline loss of Rb function in quiescent cells. **a**, Incorporation of BrdU into quiescent $Rb^{-/-}$ ($n = 4$) and $cRb^{lox/lox}$ ($n = 6$) MEFs infected with Ad (filled bars) or Ad-Cre (open bars) and incubated with BrdU for 24 hours at 3 days after infection. Under the same conditions, more than 90% incorporation occurs in proliferating cells (not shown). **b**, Immunoblot analysis of E2F target genes and cell cycle inhibitors in two representative independent quiescent $Rb^{-/-}$ (lanes 1 and 2) and $cRb^{lox/lox}$ (lanes 3 and 4) MEF populations before infection with Ad or Ad-Cre (asterisk, non-specific protein bands). **c**, Left, immunoblot analysis of pRB family members in quiescent $Rb^{-/-}$ (lanes 1 and 2) and $cRb^{lox/lox}$ (lanes 3 and 4) MEF populations before infection with adenovirus. Right, comparison of p107 levels in proliferating and quiescent $Rb^{-/-}$ (lanes 1 and 2) and $cRb^{lox/lox}$ (lanes 3 and 4) MEF populations before infection with adenovirus. Levels of p107 are lower in quiescent $Rb^{-/-}$ cells than in proliferating $Rb^{-/-}$ cells (not shown); different exposures were used to show the relative

difference between the two conditions. A non-specific protein immunoreactive band confirmed equal loading (not shown). **d**, Northern blot analysis of p107 RNA in quiescent $Rb^{-/-}$ (lanes 1–3) and $cRb^{lox/lox}$ (lanes 4–6) MEF populations using a probe encompassing p107 exon 8. Levels of the control ARPP P0 RNA are similar in the two populations. **e**, Western blot analysis of p107 in $Rb^{-/-}$ MEFs stably infected with a control retrovirus (–) or a retrovirus producing siRNA molecules against p107 (si-p107; 19-mer 5'-GATGTTGTCAGTGAGAGAG-3') (asterisk, non-specific protein bands). **f**, Incorporation of BrdU into fully quiescent GFP-positive $Rb^{-/-}$ MEFs ($n = 3$) infected with a control GFP lentivirus (si) or a GFP lentivirus expressing siRNA molecules against p107 (si-p107) and incubated with BrdU for 24 h at 3 days after infection. **g**, Time course of BrdU incorporation into quiescent $cRb^{lox/lox}$ MEFs after infection with Ad-Cre. Cells were infected with Ad or Ad-Cre 3, 6 or 10–12 days before they were considered quiescent in our initial protocol.

senescence in culture^{4,13,14}. However, MEFs with targeted deletion of all three *Rb* family members do not^{4,13}. These data could indicate that the different members of the pRB family are equally capable of inducing and maintaining the senescent state, another example of functional overlap and functional compensation within this gene family. However, this compensation is difficult to reconcile with the fact that, whereas *RB* mutations are common in human cancer, *p107* and *p130* mutations occur rarely, if ever¹⁵.

We therefore decided to study the effect of acute *Rb* mutation in senescent cells. We found that p107 concentrations were higher in senescent *Rb*^{-/-} MEFs than in senescent wild-type MEFs (Fig. 3a), indicating that p107 might also compensate for a loss of pRB in this system. *cRb*^{lox/lox} MEFs were cultured following 3T3 or a 3T2 protocol¹⁶; they were infected with Ad or Ad-Cre when they reached senescence and stopped accumulating (passage 8–12 of a 3T2 assay). Because populations of senescent cells are not fully synchronized and can contain a subpopulation of immortalized clones, we excluded cultures in which more than 15% of cells were BrdU-positive. The selected *cRb*^{lox/lox} MEF populations were infected with Ad or Ad-Cre, and 3 days later (to allow pre-existing pRB protein to be degraded; see Fig. 1c), BrdU was added for 24 h before fixation and staining. Under these conditions, senescent Ad-infected cells displayed 10–20% BrdU-positive cells. In contrast, after infection with Ad-Cre, 50% of these cells became BrdU-positive (Fig. 3b),

indicating that the cell cycle arrest associated with senescence can be reversed after acute loss of *Rb*. This result was confirmed in early-passage *cRb*^{lox/lox} MEFs induced to senesce prematurely with a retroviral construct expressing high levels of oncogenic Ras (H-Ras^{V12})¹⁷ (Fig. 3c).

Human senescent cells remain blocked in G2 or die after induction of the simian virus 40 large T oncoprotein¹⁸. In contrast, 27 of 100 senescent *cRb*^{lox/lox} cells infected with Ad-Cre rounded up 3 days after addition of virus (and only 2 of 100 in Ad-infected cells), as observed with a recording period of 40 h using time-lapse videomicroscopy (Supplementary Fig. 2). These experiments confirm that senescent *cRb*^{lox/lox} cells can complete the cell cycle after an acute loss of *Rb* function.

Injection of anti-p53 antibodies has been shown to reverse the senescent state in human cells temporarily¹⁹. To determine whether dividing post-senescent *cRb*^{-/-} MEF populations could undergo multiple cell divisions, *cRb*^{lox/lox} cells were subjected to the 3T3 protocol until they reached senescence (as assessed by morphological changes and a lack of accumulation in cell number). Populations containing obvious small immortalized cells were not included for study. Selected senescent populations were then divided in two and infected with Ad or Ad-Cre. Ad-infected *cRb*^{lox/lox} cells remained senescent and did not show an increase in cell number. In contrast, 14 of 16 *cRb*^{lox/lox} MEF populations infected with Ad-Cre resumed

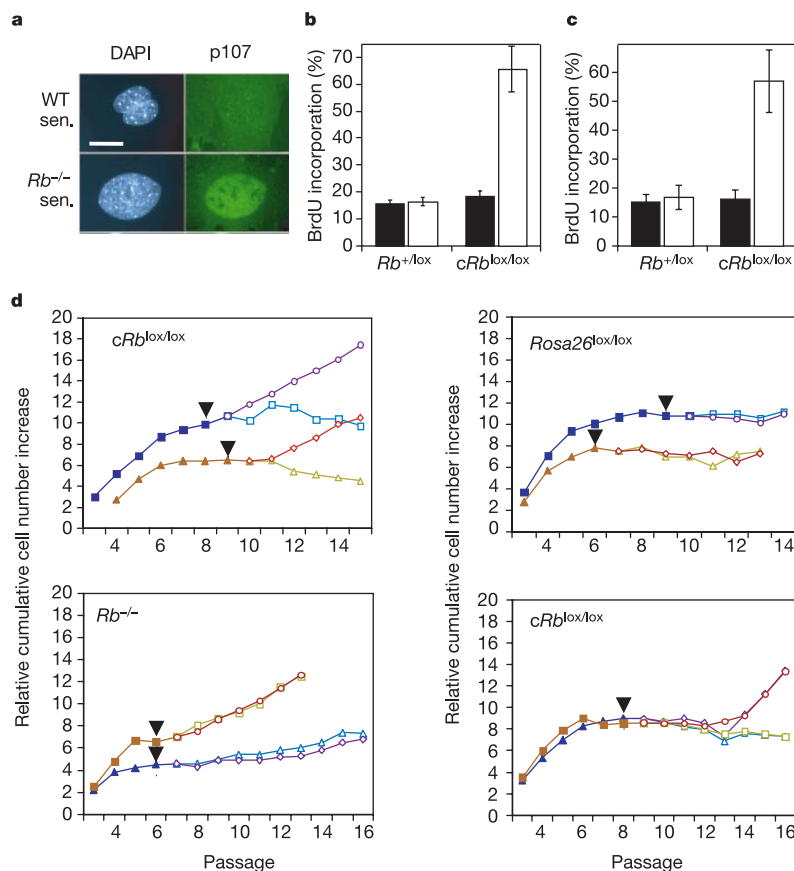


Figure 3 Reversal of senescence after acute loss of *Rb*. **a**, Immunostaining analysis of p107 levels (green) in wild-type (WT) and *Rb*^{-/-} senescent (sen.) MEFs. Staining of the DNA with DAPI (blue) shows the nucleus. The specificity of the antibody was confirmed by the absence of nuclear signal from *p107*^{-/-} MEFs (not shown). Scale bar, 50 μm.

b, c, Senescent populations of *cRb*^{lox/lox} (*n* = 6) and *cRb*^{+lox} (*n* = 2) MEFs obtained at the end of a 3T2 protocol (**b**) and prematurely senescent populations of *cRb*^{lox/lox} (*n* = 4) and *cRb*^{+lox} (*n* = 3) MEFs expressing high concentrations of oncogenic H-Ras^{V12} (**c**)

were infected with Ad (filled bars) or Ad-Cre (open bars). After 3 days, BrdU was added for 24 h before fixation, staining and counting. **d**, *cRb*^{lox/lox} (top left and bottom right panels), *Rb*^{-/-} (bottom left) and *Rosa26*^{lox/lox} (top right) MEFs were grown following a 3T3 protocol (two representative examples are shown, filled squares and filled triangles). At senescence (arrow), MEFs were infected with Ad (top panels and bottom left) or LV-lacZ (bottom right, open squares and open triangles), or Ad-Cre (top panels and bottom left) or LV-Cre (bottom right, open circles and open diamonds).

proliferation and showed clear cell accumulation within three or four passages (Fig. 3d). PCR and immunoblot analysis showed that these growing cells had deleted *Rb* exon 3 and did not express pRB (data not shown). As controls for this experiment, we have shown that wild-type MEFs with *loxP* sites in a non-relevant gene (*Rosa26*; $n = 3$) and germline *Rb*^{-/-} MEFs ($n = 3$) did not proliferate further after adenovirus infection (Fig. 3d). In addition, senescent *cRb*^{lox/lox} cells infected with a lentiviral vector expressing *Cre* (LV-*Cre*)²⁰ re-entered the cell cycle and resumed proliferation ($n = 3$), whereas cells infected with a control virus (LV-lacZ) remained senescent (Fig. 3d). The slower accumulation of cells after infection with LV-*Cre* might be due to the fact that only ~50–60% of the cells were infected by the lentiviral vector as assessed by β -galactosidase staining. Finally, post-senescent cycling *cRb*^{-/-} cells remained sensitive to the reintroduction of *Rb* (data not shown), indicating that cells had not undergone major mutations after the reversal of senescence.

Further analysis of post-senescent *cRb*^{-/-} cells showed that other markers associated with senescence (cell size, plasminogen activator inhibitor-1 (PAI-1)) were also dependent on *Rb* function, indicating that pRB might control multiple downstream aspects of the senescence programme (Supplementary Fig. 3). The p53 pathway was still intact, and levels of p19^{ARF}, p16^{INK4a} and p21^{CIP1} also remained constant or increased in cycling post-senescent *cRb*^{-/-} cells in comparison with pre-senescent and senescent cells (Supplementary Fig. 3). Levels of p107 also increased in senescent *cRb*^{lox/lox} cells 4 days after the loss of *Rb* and in further passages (data not shown). It is difficult to know whether the increase in p19^{ARF} and p107 levels occurred because more cells were in active cycle or whether it was a direct result of a compensatory mechanism. One might expect that increased levels in cell cycle inhibitors in post-senescent *cRb*^{-/-} cells would ultimately lead to the re-establishment of senescence, even though cycling cells in 10% serum might be less sensitive than quiescent or senescent cells to their growth-inhibitory action. Indeed, 2 of 14 *cRb*^{-/-} MEF populations re-entered senescence at passage 18–20 (data not shown). The long-term fate of the remaining populations was uncertain owing to the emergence of immortalized clones. The results of these experiments strongly suggest that senescent MEFs can re-enter the cell cycle, divide and proliferate after the acute loss of *Rb*, even in cells with wild-type *p107* and *p130*. Similar results have recently been obtained after loss of *p53* function in MEFs²¹; taken together, these results show that the senescence-induced cell cycle arrest is not an irreversible process.

By studying two cellular contexts, quiescence and senescence, we describe functions for the *Rb* tumour suppressor gene that were not apparent from previous analysis with cells derived from germline *Rb*^{-/-} embryos. These results have important implications for the interpretation of knockout studies generally, and they might also inform the design of mouse models of cancer involving *RB* and other cancer-associated genes. The most likely explanation for the relative normality of *Rb*^{-/-} cells in culture is functional compensation by the family members p107 and p130. Thus, *Rb*^{-/-} cells might upregulate p107 to help establish and maintain quiescent and senescent cell cycle arrest. However, in a situation of functional overlap and potential functional compensation, it is difficult to explain why the loss of a given gene function (for example, *Rb*) during tumorigenesis would have a significant effect. Our data might help to resolve this apparent conflict. By detecting phenotypic differences between conditional and acute loss of *Rb* function, we conclude that functional compensation is context dependent. Although such compensation clearly can occur, it might require a more prolonged absence of *Rb* function and a more complex resetting of the cell cycle machinery. Therefore, simple, acute loss of *Rb* can by itself stimulate quiescent and senescent cells to re-enter the cell cycle. The additional rounds of cell division that would ensue from an acute loss of *Rb* function could be crucial in the control of tumour initiation and progression. Given these

considerations, cancer modelling strategies that employ a conditional loss of *Rb*^{22,23}, other tumour suppressor genes and even gain-of-function mutations^{2,3} might be important in accurately recapitulating the effects of spontaneous mutations in human cancer. □

Methods

Embryonic stem cells and gene targeting

Two embryonic stem cell clones were introduced into the germline of mice, one in which the puromycin selection cassette was left in the intron downstream of exon 3 (ref. 4), and another in which this cassette was excised after the transient expression of *Cre* in embryonic stem cells. Cells produced from mice derived from the two types of clones gave identical results.

Culture of MEFs

All experiments were performed with MEFs prepared from embryos from at least two different litters in a mixed 129Sv/J;C57Bl/6 background. MEFs with *loxP* sites flanking a transcriptional stop cassette upstream of the *lacZ* gene in the *Rosa26* locus (*Rosa26*^{lox/lox}) were generated from mice provided by Dr S. Orkin. Serial 3T2 and 3T3 cultivation was conducted as described^{1,16}.

MEFs were infected with high-titre retroviral or lentiviral stocks produced by the transient transfection of 293T cells^{4,9}. After infection with the pSIRIPP retrovirus allowing the expression of siRNA molecules (a derivative of pSuper²⁴ and pQCXN, provided by S. Lessnick), MEFs were selected with 2 $\mu\text{g ml}^{-1}$ puromycin. After infection with the pWZL-H-Ras^{V12} retrovirus (a gift from D. Tuveson), MEFs were selected with 75 $\mu\text{g ml}^{-1}$ hygromycin B.

Adenoviral stocks were purchased from the Gene Transfer Vector Core facility of University of Iowa College of Medicine. About 100 plaque-forming units of virus were used per cell, and the medium was changed 1 day after infection.

Cell cycle, cell size and cell death assays

These assays were performed as described⁴. DNA synthesis was assayed by the incorporation of BrdU (3 $\mu\text{g ml}^{-1}$; Sigma) for 24 h with an anti-BrdU antibody coupled with fluorescein isothiocyanate (Roche; dilution 1:50) or a rat anti-BrdU antibody (Becton Dickinson; dilution 1:500). At least 500 cells were counted for each experiment.

Immunodetection

Western blot analysis of whole-cell extracts for cell cycle proteins and PAI-1 (Transduction Laboratories, P30620; dilution 1:1000) was performed as described⁴. Immunofluorescence analysis of p107 was performed on paraformaldehyde-fixed cells by using the SD15 monoclonal antibody (Neomarkers; dilution 1:20).

RNA analysis

RNA preparation and northern blot analysis were performed as described⁴. *p107* and *p130* oligonucleotides designed for real-time PCR span several exons at the 3' end of the complementary DNAs; sequences and protocol are available upon request.

Received 19 March; accepted 1 May 2003; doi:10.1038/nature01764.

- Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
- Van Dyke, T. & Jacks, T. Cancer modelling in the modern era: Progress and challenges. *Cell* **108**, 135–144 (2002).
- Jonkers, J. & Berns, A. Conditional mouse models of sporadic cancer. *Nature Rev. Cancer* **2**, 251–265 (2002).
- Sage, J. *et al.* Targeted disruption of the three *Rb*-related genes leads to loss of G₁ control and immortalization. *Genes Dev.* **14**, 3037–3050 (2000).
- MacPherson, D. *et al.* Conditional mutation of *Rb* causes cell cycle defects without apoptosis in the central nervous system. *Mol. Cell. Biol.* **23**, 1044–1053 (2003).
- Hurford, R. K. Jr, Cobrinik, D., Lee, M. H. & Dyson, N. pRB and p107/p130 are required for the regulated expression of different sets of E2F responsive genes. *Genes Dev.* **11**, 1447–1463 (1997).
- Lomazzi, M., Moroni, M. C., Jensen, M. R., Frittoli, E. & Helin, K. Suppression of the p53- or pRB-mediated G₁ checkpoint is required for E2F-induced S-phase entry. *Nature Genet.* **31**, 190–194 (2002).
- Trimarchi, J. M. & Lees, J. A. Sibling rivalry in the E2F family. *Nature Rev. Mol. Cell Biol.* **3**, 11–20 (2002).
- Rubinson, D. A. *et al.* A lentivirus-based system to functionally silence genes in primary mammalian cells, stem cells and transgenic mice by RNA interference. *Nature Genet.* **33**, 401–406 (2003).
- Hayflick, L. & Moorhead, P. S. The serial cultivation of human diploid cell strains. *Exp. Cell Res.* **25**, 585–621 (1961).
- Campisi, J. Cellular senescence as a tumour-suppressor mechanism. *Trends Cell Biol.* **11**, S27–S31 (2001).
- Schmitt, C. A. *et al.* A senescence program controlled by p53 and p16^{INK4a} contributes to the outcome of cancer therapy. *Cell* **109**, 335–346 (2002).
- Dannenberg, J. H., van Rossum, A., Schuijff, L. & te Riele, H. Ablation of the retinoblastoma gene family deregulates G₁ control causing immortalization and increased cell turnover under growth-restricting conditions. *Genes Dev.* **14**, 3051–3064 (2000).
- Carneiro, C. *et al.* p27 deficiency desensitizes *Rb*^{-/-} cells to signals that trigger apoptosis during pituitary tumour development. *Oncogene* **22**, 361–369 (2003).
- Classon, M. & Harlow, E. The retinoblastoma tumour suppressor in development and cancer. *Nature Rev. Cancer* **2**, 910–917 (2002).
- Todaro, G. & Green, H. Quantitative studies of the growth of mouse embryo cells in culture and their development into established lines. *J. Cell Biol.* **17**, 299–313 (1963).
- Serrano, M., Lin, A. W., McCurrach, M. E., Beach, D. & Lowe, S. W. Oncogenic ras provokes premature

- cell senescence associated with accumulation of p53 and p16INK4a. *Cell* **88**, 593–602 (1997).
18. Jat, P. S. & Sharp, P. A. Cell lines established by a temperature-sensitive simian virus 40 large-T-antigen gene are growth restricted at the nonpermissive temperature. *Mol. Cell. Biol.* **9**, 1672–1681 (1989).
 19. Gire, V. & Wynford-Thomas, D. Reinitiation of DNA synthesis and cell division in senescent human fibroblasts by microinjection of anti-p53 antibodies. *Mol. Cell. Biol.* **18**, 1611–1621 (1998).
 20. Pfeifer, A., Brandon, E. P., Kootstra, N., Gage, F. H. & Verma, I. M. Delivery of the Cre recombinase by a self-deleting lentiviral vector: Efficient gene targeting *in vivo*. *Proc. Natl Acad. Sci. USA* **98**, 11450–11455 (2001).
 21. Dirac, A. M. & Bernards, R. Reversal of senescence in mouse fibroblasts through lentiviral suppression of p53. *J. Biol. Chem.* **278**, 11731–11734 (2003).
 22. Vooijs, M., van der Valk, M., te Riele, H. & Berns, A. Flp-mediated tissue-specific inactivation of the retinoblastoma tumour suppressor gene in the mouse. *Oncogene* **17**, 1–12 (1998).
 23. Marino, S., Vooijs, M., van der Gulden, H., Jonkers, J. & Berns, A. Induction of medulloblastomas in p53-null mutant mice by somatic inactivation of Rb in the external granular layer cells of the cerebellum. *Genes Dev.* **14**, 994–1004 (2000).
 24. Brummelkamp, T. R., Bernards, R. & Agami, R. A system for stable expression of short interfering RNAs in mammalian cells. *Science* **296**, 550–553 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank R. Jaenisch, S. Lowe and B. Kennedy for various useful reagents; D. Tuveson for helpful discussions; A. Brunet, K. Johnson and M. McLaughlin for critical reading of the manuscript; and members of F. Gertler's laboratory for help with the videomicroscopy experiments. This work was supported by funding from the Human Frontier Science Program, the Medical Foundation and the Merck/MIT postdoctoral fellowship program (J.S.), and from the National Cancer Institute and the Howard Hughes Medical Institute (T.J.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.J. (tjacks@mit.edu).

Xpd/Ercc2 regulates CAK activity and mitotic progression

Jian Chen, Stéphane Larochelle*, Xiaoming Li & Beat Suter

Department of Biology, McGill University, Montreal, PQ, Canada H3A 1B1

General transcription factor IIH (TFIIH) consists of nine subunits: cyclin-dependent kinase 7 (Cdk7), cyclin H and MAT1 (forming the Cdk-activating-kinase or CAK complex), the two helicases Xpb/Hay and Xpd, and p34, p44, p52 and p62 (refs 1–3). As the kinase subunit of TFIIH, Cdk7 participates in basal transcription by phosphorylating the carboxy-terminal domain of the largest subunit of RNA polymerase II^{1,4,5}. As part of CAK, Cdk7 also phosphorylates other Cdks, an essential step for their activation^{6–9}. Here we show that the *Drosophila* TFIIH component Xpd negatively regulates the cell cycle function of Cdk7, the CAK activity. Excess Xpd titrates CAK activity, resulting in decreased Cdk T-loop phosphorylation, mitotic defects and lethality, whereas a decrease in Xpd results in increased CAK activity and cell proliferation. Moreover, Xpd is downregulated at the beginning of mitosis when Cdk1, a cell cycle target of Cdk7, is most active. Downregulation of Xpd thus seems to contribute to the upregulation of mitotic CAK activity and to regulate mitotic progression positively. Simultaneously, the downregulation of Xpd might be a major mechanism of mitotic silencing of basal transcription.

With its dual role, Cdk7 might function as a critical link between the cell cycle and transcriptional programmes^{1,6,10}. To identify genes involved in directing Cdk7 activity towards transcription or cell-cycle function in a multicellular organism, we screened for dominant suppressors and enhancers of a hypomorphic *Drosophila* *cdk7^{ts1}* mutant that has a cell cycle but no transcriptional defect.

We found that decreasing *xpd/Ercc2* (xeroderma pigmentosum disorder group D/excision repair cross-complementing) strongly suppresses the hypomorphic *cdk7^{ts1}* cell-cycle defect (Supplementary Information).

The first 13 mitotic cycles of the *Drosophila* embryo are synchronized syncytial cycles that do not require transcription and are composed of only the S and M phases. During cycle 14, cell divisions become asynchronous¹¹ and zygotic transcription increases strongly. On such embryos the anti-Xpd antibody produces a highly cell-cycle-dependent staining pattern (Fig. 1a–e). Figure 1a–c shows asynchronous cells of a wild-type embryo at cycle 14. In interphase and prophase cells, a strong Xpd signal is found. In cells at metaphase, anaphase and telophase, the signal drops to almost background levels. A similar pattern was observed in embryos at cycle 15 (not shown), and in embryos at cycle 13 the Xpd signal is significantly lower during metaphase than in interphase (Fig. 1d, e). In contrast, immunostaining with antibodies against Xpb/Hay did not show this cell-cycle dependence (not shown).

To find out whether the drop in Xpd signal is caused by the downregulation of Xpd, Schneider cells were synchronized at G2/M and aliquots of the cells were collected at the indicated time points after release from the block (Fig. 1f). Starting at 120 min, the percentage of cells in G1 increases at the expense of G2/M cells, indicating that G2/M arrested cells have started to progress through mitosis. Also at 120 min the Xpd concentrations start to drop. At 240 min, Xpd has reached its lowest concentration, coinciding with the rapid downregulation of cyclin B and cyclin A. Together with the embryonic *in situ* staining data, this result indicates that the Xpd polypeptide is downregulated during prometaphase.

To test whether low Xpd concentrations are required for normal CAK activity and for cells to go through mitosis, we overexpressed Xpd in 2–4-h-old embryos. Embryos containing two copies of the *hsp-xpd* transgene were heat shocked for 30 min at 36 °C, and CAK activity was measured after recovery. Extracts from transgenic embryos had significantly less CAK activity than non-transgenic controls (Fig. 2a, b). To find out whether these differences in CAK activity are physiologically relevant, Cdk1 phosphorylation was analysed after pulse induction of Xpd. *Drosophila* Cdk1 polypeptide can be resolved into three to four isoforms, with the fastest migrating isoform, isoform 1, being the Thr 161-phosphorylated Cdk1 (Fig. 2c)¹². In a comparable manner to non-heat-shocked wild-type flies, wild-type embryos have significant concentrations of the Thr 161-phosphorylated isoform 1 after a 60-min recovery period. However, in flies containing the *hsp-xpd* transgene, the Thr 161-phosphorylated isoform is still undetectable after 60 min of recovery. Overexpression of Xpd thus results in less phosphorylation of Cdk1 on Thr 161.

In 2–3-h-old wild-type embryos, the heat treatment caused only 10% of the embryos to show a lethal phenotype 90 min into the recovery period. However, on overexpression of *xpd*, 50% of the embryos showed lethality (Fig. 2d) as judged by large areas of DNA aggregation, irregular mitotic figures and degenerating cells (not shown). The embryonic lethality is comparable to the overall lethality during the entire life, indicating that pulse overexpression of *xpd* in the embryo has a rapid effect only on embryonic viability and no long-term effects are caused by, for instance, general transcriptional problems. This lethality was not observed with control transgenic lines containing either *hsp-xpb/hay* or an *hsp-xpd* construct that lacks the amino-terminal CAK-interaction domain¹³ (*hsp-xpdC*), indicating that decreased fly viability is indeed caused by greater Xpd concentrations (Fig. 2d).

We also tested the effect of different *xpd* dosage during the entire life cycle by using a co-culture assay with offspring from a single cross (Fig. 2e). After daily induction of *hsp-xpd* in addition to the wild-type *xpd* concentration, the viability of transgenic *hsp-xpd* lines (*CyO/+*; *+hsp-xpd*) is decreased to 14% (36 of 251) compared with the wild-type 100% (*CyO/+*; *+PrDr*). This decreased

* Present address: Cell Biology Program, Memorial Sloan-Kettering Cancer Center, New York, New York 10021, USA.

viability is partly restored to 40% (100 of 251) in the siblings that lack one copy of endogenous *xpd* (*Df(xpd)/+*; *+/-hsp-xpd*). In addition, the viability of hemizygous flies containing only one endogenous copy of *xpd* (*Df(xpd)/+*; *+/-PrDr*) is decreased to about half that of the wild type (57%, 143 of 251) under our experimental conditions with heat shock stress. This indicates that the concentration of Xpd serves critical cellular functions.

Loss of *cdk7* activity in *Caenorhabditis elegans* embryos leads to cell-cycle arrest in M phase and interphase¹⁴ and to cell-division defects in *Drosophila*⁹. To test whether Xpd overexpression results in a similar phenotype, we stained embryos with the antibody against the mitotic marker phospho-histone H3 (PH3) at 45 min into the recovery period. At this time, embryos have restarted their cell cycle and *xpd*-overexpressing embryos have not yet started to die (Fig. 2d). In wild-type embryos (Fig. 3a, c), cells in each mitotic domain enter mitosis in a slightly asynchronous manner, starting from the centre and continuing to the outside of each domain¹¹. In domains 1 and 4, telophase, metaphase and prophase figures are seen concentrically in this order from the centre to the outside (Fig. 3a, c, a', c'). The embryo in Fig. 3c is slightly (a few minutes) older than that in Fig. 3a, and a wider central area devoid of PH3 staining is seen in the former, indicating that these cells have exited from mitosis (arrows in Fig. 3a', c'). Because cells within one mitotic domain are normally not completely synchronized, the

appearance of larger patches of cells in the same mitotic phase indicates a cell-cycle arrest. In mitotic domain 1 of embryos from *xpd*-overexpressing lines, cells undergo metaphase arrest and there are many additional metaphase figures (arrowheads in Fig. 3b'; compare with Fig. 3a', c'). Similar metaphase arrests are seen in domains 1 and 4 of the *xpd*-overexpressing embryo (compare Fig. 3d with Fig. 3c). In addition, we observe prophase arrest in domain 4 of *xpd*-overexpressing embryos (compare Fig. 3b'' with Fig. 3a, c). Various metaphase and prophase figures are seen in domains 5 and 6 in wild-type embryos (Fig. 3a, c). However, the mitotic PH3 signal is not seen (Fig. 3b) or is less frequent (Fig. 3d) in domains 5 and 6 of *xpd*-overexpressing embryos, indicating that cells in these domains are unable to enter M phase, probably arresting at G2/M. This difference is not due to the younger age of the *xpd*-overexpressing embryos because the extent of their mitotic domains 1 and 4 indicates that they are at least as old as the wild-type embryos in Fig. 3a, c. Overexpression of *xpd* therefore causes several distinct cell-cycle arrests at G2/M, in prophase and in metaphase. Furthermore, the fact that morphogenesis progresses normally shows that under our experimental conditions *xpd* overexpression does not block all basal transcription. Consistent with this, we find that cyclin

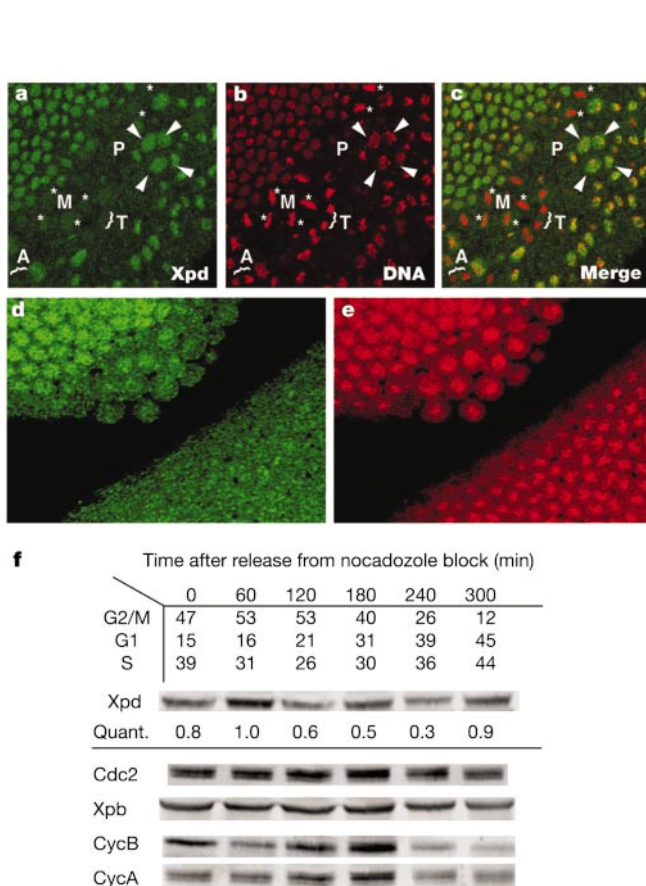


Figure 1 Cell-cycle dependence of Xpd. **a-c**, Confocal images of cycle-14 embryos stained for Xpd (**a**), DNA (**b**) and both (**c**). M and asterisks show metaphase figures; P and arrowheads show prophases; A and T with braces show anaphases and telophases, respectively. **d, e**, Cycle-13 embryos stained for Xpd (**d**) and DNA (**e**). The weaker Xpd signal coincides with condensed chromosomes. **f**, Western blot showing Xpd concentrations in S2 cells at the indicated time points after release from the nocodazole block. Tabulated numbers indicate the percentages of cells in the cell-cycle phases given. Quant., quantified signals were normalized to Cdc2 (Cdc2), with the highest value set to 1.0.

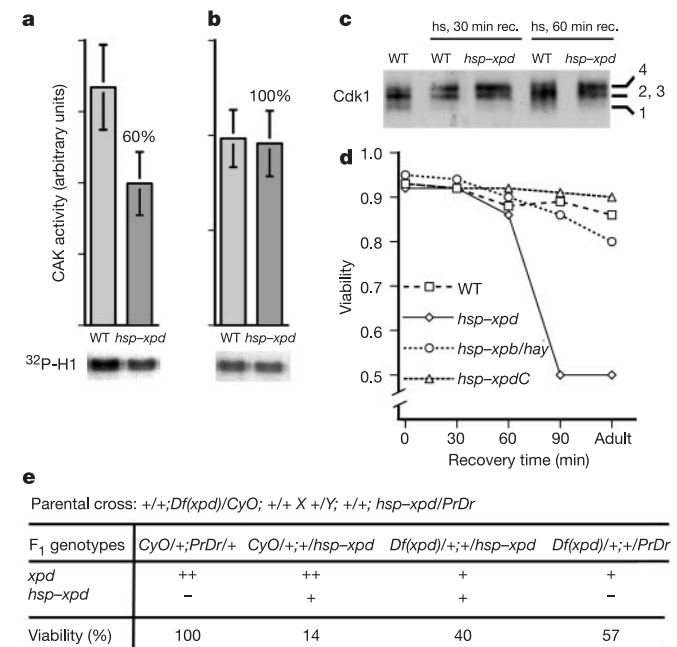


Figure 2 Reduced CAK activity and viability in embryos overexpressing *xpd*. **a, b**, CAK activity after (**a**) and without (**b**) *hsp-xpd* induction. The autoradiography signal of ³²P-histone H1 (³²P-H1) was quantified and is displayed on the y-axis. WT, wild type. **c**, Thr 161-phosphorylated Cdk1 (isoform 1) is still decreased after a 60-min recovery (rec.) in *hsp-xpd* embryos, but not in wild-type embryos. Isoform 1, Thr 161-phosphorylated Cdk1; isoform 2, unphosphorylated Cdk1; isoform 3, Tyr 15-phosphorylated Cdk1; isoform 4, Thr 14, Tyr 15-phosphorylated Cdk1 (ref. 12). **d**, Embryonic viability was evaluated by staining mitotic spindles (anti- α -tubulin) and DNA (Hoechst 33258). Previous reports have described that heat shock causes a transient interphase arrest of nuclei in syncytial embryos²⁰ and cellularized embryos²¹. Most embryos show mitotic spindles and condensed DNA 30 min after heat shock (hs) treatment. They have recovered from the heat shock and re-entered mitosis. This recovery time is close to that described by others²¹. After 90 min of recovery, *xpd*-overexpressing embryos (but not controls) show high lethality. **e**, Viability of flies at four different *xpd* concentrations. *CyO/+*; *+/PrDr* flies have wild-type *xpd* concentrations and their viability serves as a control (100%). Each plus sign refers to one copy of the endogenous *xpd* or the *hsp-xpd* transgene.

A and B concentrations are normal as assessed by western blot analysis (not shown).

We also induced *xpd* overexpression in embryos at cycle 13 and aged them to cycle 14 before fixing and staining them with anti- α -

tubulin antibodies to detect their mitotic spindles (Fig. 3e, e', f, f'). Both wild-type and *hsp-xpd* transgenic embryos are in cycle 14, as shown by the appearance of cycle 14 mitotic domains and by their morphogenetic stage. However, wild-type embryos (Fig. 3e) contain about twice as many cells as the *xpd*-overexpressing embryos (Fig. 3f). In the wild type, 23 cells could be counted along the cephalic furrow and 22 from the anterior of the head to the cephalic furrow. In *xpd*-overexpressing embryos, only 16 and 15 cells are counted along the corresponding lines. This suggests that *xpd* overexpression in cycle 13 causes these cells to skip mitotic phase 13 and to continue with cycle 14. Supporting this interpretation is the fact that there is no sign of cell-cycle arrest or cell lethality and that the cells in *hsp-xpd* embryos are larger (Fig. 3f, e, f'). The absence of a cell cycle arrest on the overexpression of *xpd* during cycle 13 might have to do with the lack of an appropriate cell-cycle checkpoint in this syncytial cycle.

Cdk7 localizes to the nucleus in most differentiated cells. However, in young embryos its localization pattern is somewhat different. In wild-type embryos, mitotic cells of mitotic domain 1 have an abundant Cdk7 signal throughout the cytoplasm and the nucleoplasm (Fig. 3g, h). Interphase cells, in contrast, have most Cdk7 signal in the cytoplasm, whereas the nuclear Cdk7 signal is concentrated on the DNA (Fig. 3g, h). We compared the accumulation of Cdk7 and Xpd with or without Xpd overexpression. In

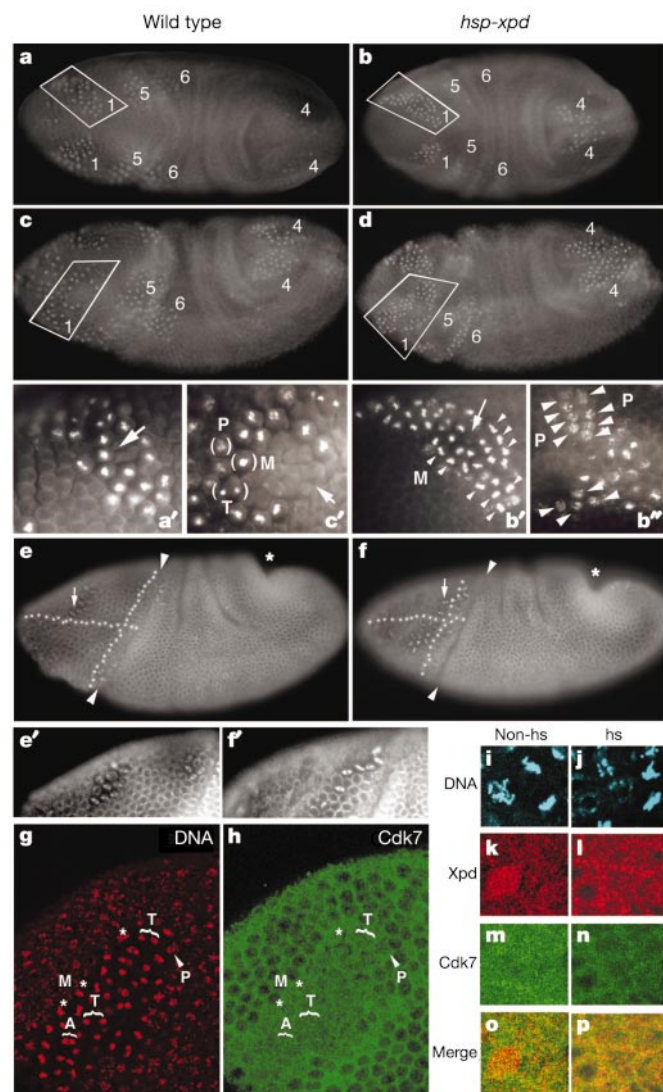


Figure 3 Cellular effects of *xpd* overexpression. **a–d**, Cell-cycle arrest in cycle-14 embryos shown by anti-PH3 staining. Wild-type embryos (**a, c**) and *hsp-xpd* embryos (**b, d**) are at about the same gastrulation stage. **a, b**, Ventral views; **c, d**, ventrolateral views. The numbers 1, 4, 5 and 6 indicate the respective mitotic domains. **a', b', c'**, Magnified domain 1 (boxed region) of the embryos in **a, b** and **c**, respectively. **b''**, Magnified domain 4 of the embryo in **b**. P, prophase figures; M, metaphase; T, telophase. Arrows point to the central region of the mitotic domains, where cells have exited from mitosis and lost the PH3 signal. Arrowheads in **b'** and **b''** point to metaphase and prophase figures, respectively. **e, f**, Overexpression of *xpd* in cycle-13 embryos results in fewer cell divisions. Anti- α -tubulin stainings of cycle-14 embryos are shown in a lateral view. Mitotic spindles (arrows) are seen in both embryos. Wild-type embryo (**e**) and *hsp-xpd*-expressing embryo (**f**) undergo gastrulation and are at the same developmental stage. Asterisk, expanding germ band. Two arrowheads indicate the position of the cephalic furrow. The white dotted lines parallel and perpendicular to the cephalic furrow aid in counting the cells along the anteroposterior and dorsoventral axes. **e', f'**, Magnified view of the head regions of **e** and **f**, respectively. **g, h**, Cdk7 localization is dependent on the cell cycle in mitotic cycle-14 embryos. M and asterisk, metaphase figures; P, prophase; A, anaphase; T, telophase. **i–p**, Xpd overexpression results in mislocalization of Xpd and Cdk7. Confocal images from mitotic cycle-14 embryos show DNA staining in blue, Xpd signal in red and Cdk7 signal in green. **i, k, m, o**, Wild-type levels of *xpd* (non-heat-shock control); **j, l, n, p**, overexpression of Xpd was induced by heat shock (hs).

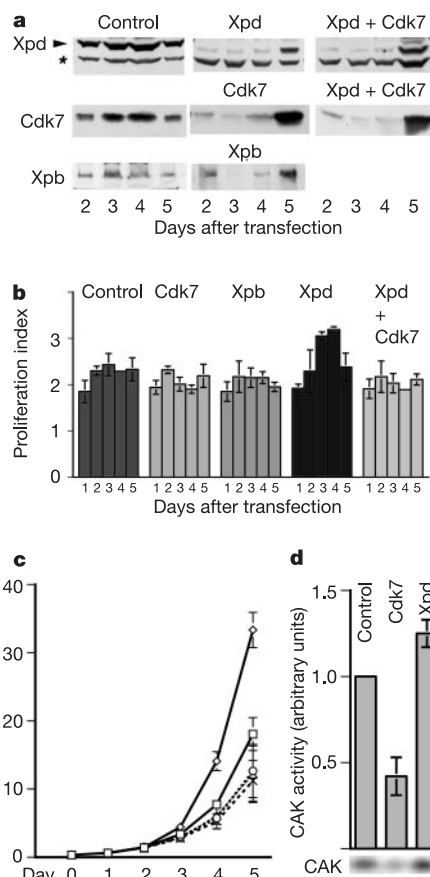


Figure 4 Decreased Xpd concentration is correlated with increased CAK activity and cell proliferation in S2 cells. **a**, Western blots show protein concentrations of Xpd, Cdk7 and Xpb/Hay 2–5 days after RNAi treatment. Protein concentrations return to normal on day 5. Asterisk, a non-specific cross-reactive band serves as an additional loading control. **b**, Proliferation index (number of cells on day of count divided by number of cells on previous day) after RNAi treatment. **c**, Numbers of viable cells during the course of the experiment. Squares, control; diamonds, Xpd; circles, Cdk7; triangles, Xpb; crosses, Xpd + Cdk7. **d**, Average CAK activity on the fourth day after RNAi treatment. Control was set at 1.0.

control embryos with wild-type levels of *xpd*, strong Xpd staining is seen in the nucleus in prophase, but not metaphase, cells (Fig. 3i, k; see also Fig. 1a–c). During this period the Cdk7 signal becomes uniformly distributed (Fig. 3m, o; see also Fig. 3h). However, in *xpd*-overexpressing embryos, the staining patterns of Cdk7 and Xpd are altered and signals localize together in the different cell-cycle phases (Fig. 3j, l, n, p). Our data indicate that Xpd overexpression alters the overall distribution of Xpd and causes the mislocalization of mitotic Cdk7.

We next considered whether decreasing the Xpd concentration affects CAK activity and cell proliferation in the opposite way to its overexpression. For this we used RNA interference (RNAi) in *Drosophila* S2 cells¹⁵. As control we also performed RNAi with *xpb/hay* and *cdk7*. Xpd, Cdk7 and Xpb/Hay concentrations are significantly decreased from day 2 to day 4 after RNAi, but they are back to normal on day 5 (Fig. 4a). Concomitantly with the decrease in Xpd concentration, increases in both cell proliferation and total cellular CAK activity are observed (Fig. 4b–d). The decrease in Cdk7 concentration results in a decrease in total cellular CAK activity (Fig. 4d), with little effect on cell proliferation (Fig. 4b, c), whereas the decrease in Xpb has no significant effect on CAK activity but slightly decreases the rate of proliferation (Fig. 4b–d). By day 5, the number of cells in the *xpd* RNAi-treated sample is double that of the control (Fig. 4c). This increase is explained by an observed increase in the rate of cell proliferation when Xpd concentration is low on days 3 and 4 after RNAi (Fig. 4b, c). On day 5, Xpd concentration is back to normal with a concomitant return to a normal rate of proliferation (Fig. 4b). These observations are in agreement with the observed effect of overexpressing Xpd in embryos, and indicate that there is a direct link between the rate of cell proliferation and the concentration of Xpd protein in *Drosophila* cells. Moreover, this enhancement of the proliferation rate is dependent on the presence of normal amounts of Cdk7 protein, as demonstrated by the fact that decreasing the amount of Cdk7 suppresses the effect of low Xpd concentration on cell proliferation (Fig. 4b).

Xpd is downregulated at mitosis, and lower Xpd concentrations increase CAK activity and promote cell-cycle progression. In contrast, Xpd overexpression can result in a decrease in total CAK activity, decreased Cdk1 T-loop phosphorylation and delayed or blocked entry into mitosis. A possible mechanism is that a high concentration of Xpd favours the incorporation of CAK into the TFIIF complex, thereby decreasing the amount of free CAK available to activate Cdk1. In eukaryotic cells, mitosis is accompanied by a global repression of transcription, caused almost exclusively (at least 90%) by the direct inactivation of components of the transcription machinery¹⁶. TFIIF is the major target of this mitotic repression, and TFIIF-associated carboxy-terminal domain (CTD) kinase activity and transcription activity are specifically repressed at mitosis^{17,18}. Xpd downregulation and general transcription silencing both occur during prophase (Fig. 1 and ref. 16, and unpublished observations). This indicates that the downregulation of Xpd might cause dissociation of the CAK complex from holo-TFIIF, causing TFIIF to lose its CTD kinase activity concurrently with the loss of Xpd. Both losses might then contribute to the shutting-down of general transcription. Simultaneously, Xpd degradation might set the trimeric CAK complex free to act as a cell-cycle promoter and to phosphorylate its cell-cycle target Cdk1. In this model, the abundance of Xpd acts as a critical regulator of both cell-cycle and basal transcription by determining the distribution of Cdk7 between its CAK and CTD kinase forms. Here we have provided tangible evidence for such a coupling between the cell cycle and the regulation of basal transcription through Xpd and Cdk7. Interactions between CAK and subunits of the proteasome machinery have previously been described¹³ and it will be interesting to discover whether these components actually regulate Xpd concentration. □

Methods

Fly culture conditions and crosses

Drosophila was cultured on standard cornmeal food at 23–25 °C unless specified otherwise. The temperature-sensitive alleles were grown at 18 °C. *cdk7*^{null} is *w* *Df(1)JB254 P[w⁺snf⁺, dhd⁺]*. The *cdk7*^{ts1} flies are *cdk7*^{null}/*cdk7*^{ts1}; +/+; *P[w⁺cdk7^{P140S}]/P[w⁺cdk7^{P140S}]* (ref. 9). *cdk7*^{S164N/T170A} was described elsewhere¹⁹.

To test whether overexpression of *xpd* causes lethality (Fig. 2e), either *Df(2R)AA21/CyO*; +/+ or *Df(2R)P113/CyO*; +/+ females were crossed to +/+; *P[w⁺hsp-xpd]/PrDr* males. Embryos were collected for 4–8 h in bottles. To induce *xpd* overexpression, the bottles were kept in an incubator that cycled between 25 °C (22 h) and 37 °C (1 h, and 1 h slope time) until adult flies hatched.

Cell culture, synchronization and flow cytometry

Drosophila S2 cell lines were cultured at 27 °C in Schneider's *Drosophila* medium, 10% FBS, 50 µg ml⁻¹ penicillin–streptomycin, 1 mM l-glutamine (all from Gibco). Cells were first synchronized at G1/S with a double thymidine block (2 mM, two 16-h treatments with a 10-h interval) followed by a G2/M block (16 h) with 0.5 µg ml⁻¹ nocodazole (Sigma). For flow cytometry analysis, cells were washed once with PBS, fixed in 75% ice-cold ethanol for 15 min and washed once with PBS. After 30 min of treatment with RNase (0.8 mg ml⁻¹; 37 °C), they were stained for 15 min with 5 µg ml⁻¹ propidium iodide (Molecular Probes) and analysed by FACScan (Becton Dickinson). Results were analysed with the ModfitLT program.

Embryo collection, Xpd induction and CAK assay

Flies were allowed to pre-lay for 1 h before egg collection. For CAK assays, embryos 2–4 h old were collected, heat-shocked for 30 min and left to recover for 45 min. Embryos used for CAK assays or immunostaining were dechorionated with 50% bleach, frozen in solid CO₂ and stored at –70 °C. CAK assays were performed as described previously⁹. For immunostaining with anti-PH3 antibody, embryos in S or G2 phase of cycle 14 were heat-shocked for 30 min and harvested 45 min thereafter. On the basis of western blot results from embryos collected after a 30-min heat shock at 36 °C, the Xpd concentration is 36% higher than in the wild type.

For induction of *xpd* overexpression in cycle 13, eggs 0–30 min old were collected and aged for 110 min to reach cycle 13. They were then heat-shocked for 30 min at 36 °C in a water bath, left to recover at 25 °C for 100 min (to reach M14), and fixed. To determine the lethal period after *xpd* induction, embryos 2–3 h old were collected and given a 30-min heat shock at 36 °C. Every 30 min an aliquot was harvested and fixed with methanol. A portion of the eggs was aged to the adult stage.

xpdC transgenic lines

The *hsp-xpdC* construct has a deletion of the 5' region from codons 6 to 441, which corresponds to codons 1–430 of its human counterpart. To generate the *hsp-xpdC* construct, the 3' part of the *xpd* cDNA was amplified by polymerase chain reaction (PCR) with sense primer 5'-ggaattccTAAATGAAAGTACTCTCCCATTTTGACATTTA-3' and antisense primer 5'-ggaattccGCAGAGCACCTTCCACATCT-3'. The resulting PCR fragment has *EcoRI* sites at both ends that allowed insertion into the *EcoRI* site of the *hsp70SCS* vector to clone it behind the *hsp70* promoter. The *hsp70-xpdC* transgene was then transferred into pCaSpeR vector by using *NotI* and *KpnI*, and transgenic lines were established.

Antibodies and immunostainings are described in Supplementary Information.

RNAi in S2 cells

RNAi experiments were performed as described previously¹⁵. PCR products were amplified from the respective cDNA clones. Primers used in the PCR reactions contain a T7 RNA polymerase promoter sequence at the 5' end (5'-TAATACGACTCACTATAGGAGG-3'), followed by sequences specific for the targeted genes. The gene-specific sequences are as follows: *cdk7* sense, 5'-TGACAAACGGAACGCTATG3', *cdk7* antisense, 5'-GCCGGTTGTGTAGCGAATA-3'; *xpb/hay* sense, 5'-GGTCAAGCTGGTCTTGAAGC-3', *xpb/hay* antisense, 5'-AGTAGTGTGGCCGCTCAATCC-3'; *xpd* sense, 5'-GACGGCCTGTTGGTGTACTT-3', *xpd* antisense, 5'-CACCTCGGTTAG GACATCGT-3'.

Received 3 February; accepted 8 May 2003; doi:10.1038/nature01746.

1. Frit, P., Bergmann, E. & Egly, J. M. Transcription factor IIF: A key player in the cellular response to DNA damage. *Biochimie* **81**, 27–38 (1999).
2. Egly, J. M. The 14th Datta Lecture. TFIIF: From transcription to clinic. *FEBS Lett.* **498**, 124–128 (2001).
3. Austin, R. J. & Biggin, M. D. Purification of the *Drosophila* RNA polymerase II general transcription factors. *Proc. Natl Acad. Sci. USA* **93**, 5788–5792 (1996).
4. Akoulitchev, S., Makela, T. P., Weinberg, R. A. & Reinberg, D. Requirement for TFIIF kinase activity in transcription by RNA polymerase II. *Nature* **377**, 557–560 (1995).
5. Makela, T. P. *et al.* A kinase-deficient transcription factor TFIIF is functional in basal and activated transcription. *Proc. Natl Acad. Sci. USA* **92**, 5174–5178 (1995).
6. Nigg, E. A. Cyclin-dependent protein kinases: Key regulators of the eukaryotic cell cycle. *BioEssays* **17**, 471–480 (1995).
7. Morgan, D. O. Cyclin-dependent kinases: Engines, clocks, and microprocessors. *Annu. Rev. Cell Dev. Biol.* **13**, 261–291 (1997).
8. Morgan, D. O. Principles of CDK regulation. *Nature* **374**, 131–134 (1995).
9. Larochelle, S., Pandur, J., Fisher, R. P., Salz, H. K. & Suter, B. Cdk7 is essential for mitosis and for in vivo Cdk-activating kinase activity. *Genes Dev.* **12**, 370–381 (1998).
10. Fisher, R. P. & Morgan, D. O. CAK in TFIIF: Crucial connection or confounding coincidence? *Biochim. Biophys. Acta* **1288**, O7–O10 (1996).

11. Foe, V. E., Odekk, G. M. & Edgar, B. A. in *The Development of Drosophila melanogaster* (eds Bate, M. & Martinez-Arias, A.) 149–300 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1993).
12. Edgar, B. A., Sprenger, F., Duronio, R. J., Leopold, P. & O'Farrell, P. H. Distinct molecular mechanism regulate cell cycle timing at successive stages of *Drosophila* embryogenesis. *Genes Dev.* **8**, 440–452 (1994).
13. Sandrock, B. & Egly, J. M. A yeast four-hybrid system identifies Cdk-activating kinase as a regulator of the XPD helicase, a subunit of transcription factor IIH. *J. Biol. Chem.* **276**, 35328–35333 (2001).
14. Wallenfang, M. R. & Seydoux, G. cdk-7 is required for mRNA transcription and cell cycle progression in *Caenorhabditis elegans* embryos. *Proc. Natl Acad. Sci. USA* **99**, 5527–5532 (2002).
15. Clemens, J. C. *et al.* Use of double-stranded RNA interference in *Drosophila* cell lines to dissect signal transduction pathways. *Proc. Natl Acad. Sci. USA* **97**, 6499–6503 (2000).
16. Gottesfeld, J. M. & Forbes, D. J. Mitotic repression of the transcriptional machinery. *Trends Biochem. Sci.* **22**, 197–202 (1997).
17. Long, J. J., Leresche, A., Kriwacki, R. W. & Gottesfeld, J. M. Repression of TFIIF transcriptional activity and TFIIF-associated cdk7 kinase activity at mitosis. *Mol. Cell. Biol.* **18**, 1467–1476 (1998).
18. Akoulitchiev, S. & Reinberg, D. The molecular mechanism of mitotic inhibition of TFIIF is mediated by phosphorylation of CDK7. *Genes Dev.* **12**, 3541–3550 (1998).
19. Larochelle, S. *et al.* T-loop phosphorylation stabilizes the CDK7-cyclin H-MAT1 complex *in vivo* and regulates its CTD kinase activity. *EMBO J.* **20**, 3749–3759 (2001).
20. Foe, V. E. & Alberts, B. M. Reversible chromosome condensation induced in *Drosophila* embryos by anoxia: visualization of interphase nuclear organization. *J. Cell Biol.* **100**, 1623–1636 (1985).
21. Maldonado-Codina, G., Llamazares, S. & Glover, D. M. Heat shock results in cell cycle delay and

synchronisation of mitotic domains in cellularised *Drosophila melanogaster* embryos. *J. Cell Sci.* **105**, 711–720 (1993).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. Zurita for *hsp-xpd* transgenic flies and the *xpd* cDNA clone; M. Fuller for anti-(Xpb/Hay) antibody, *hsp-xpb/hay* transgenic flies and *xpb/hay* cDNA clone; T. Schüpbach, W. J. Mackay and B.-Z. Shilo for fly strains; R. Tanguay for providing S2 cells; M. Miron, O. Johnstone and M. Lindor for technical assistance with the cell culture; B. Hu for microinjection of the *hsp-xpdC* construct; A. Tsang for use of the confocal microscope; A. Spurmanis for technical assistance; P. Lasko for the use of the Axioplan microscope; R. Roy for use of the PhosphorImager; and R. Roy, N. Mashrouha and M. Cavey for their critical reading of the manuscript. This work was supported by the National Cancer Institute of Canada with funds from the Canadian Cancer Society. B.S. was a research scientist of the National Cancer Institute of Canada supported by funds from the Canadian Cancer Society and is now a Canadian Institute of Health Research investigator. J.C. was a recipient of a studentship from the Fonds pour la formation de chercheurs et l'aide à la recherche (FCAR Quebec).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.S. (beat.suter@mcgill.ca or beat.suter@izb.unibe.ch).

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Global Head of Naturejobs:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)

Andy Douglas (4975)

Frank Phelan (4944)

Netherlands/ Italy/

Iberia/ Belgium:

Evelina Rubio Hakansson (4973)

Scandinavia: Silje Opstrup (4994)

France/ Switzerland:

Amelie Pequignot (4974)

Natureevents:

Paul Constant (4954)

Production Manager: Billie Franklin

To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

International

Advertising Coordinator:

Hind Berrada (4935)

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Office

Germany/ Austria:

Patrick Phelan, Odo Wulffen

Tel + 49 89 54 90 57 11/-2

Fax + 49 89 54 90 57 20

e-mail: p.phelan@nature.com

o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,

10th Floor, New York, NY 10010-1707

Tel +1 800 989 7718

Fax +1 800 989 7103

e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

US Advertising Coordinator:

Linda Adam

Japan Head Office, Tokyo

MG Ichigaya Building (5F),

19-1 Haraikatamachi,

Shinjuku-ku,

Tokyo 162-0841

Tel +81 3 3267 8751

Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami

e-mail: r.asami@naturejpn.com

naturejobs

Science rocks

In Denmark last month, scientists tried to fix a pipeline with an amplifier. Across Western Europe, the past decade has seen a decline in the number of young people going into scientific fields, and the Danes are no exception. Many are suspicious of technology, associating high-energy physics with nuclear power, chemistry with pollution, and biology with genetically modified foods.

So when the city of Roskilde approached Jens-Peter Lynov, director of the department of optics and fluid dynamics at Risø National Laboratory, to bring some of his lab's work to Europe's largest music festival, he accepted. Under the auspices of 'Musicon Valley', a riff on the region's Medicon Valley moniker, Risø applied some of its technologies to the festival. These included fibre-composite materials for stage construction, plasma treatment of materials to make them water-repellent, and biodegradable tents.

Between sets, Risø also drew a steady stream of festival-goers to a booth that highlighted these technologies, provided some basic science education, and exhibited the lab's work. "We can maybe attract some bright young people to work at our lab," says Lynov. After Roskilde, Risø will take its show on the road to other large music festivals in Scandinavia.

But it seems short-sighted to limit this alliance of rock and science to Scandinavia. The model could easily be applied to other music festivals. Lollapalooza could emphasize alternative careers, Ozzfest could garner sponsorship from alcohol and drug institutes, and Glastonbury could spotlight research on materials — especially ones involving hemp and mud. And when parents chastise their tattooed and pierced offspring en route to a festival, telling them to get a job, the kids can respond that they are on their way.

Paul Smaglik

Naturejobs editor



Contents

POSTDOCS AND STUDENTS

Getting your skills
up to speed

p234

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

The learning curve

Students starting their PhD and postdocs moving into fresh areas of research often find themselves in need of different technical skills. How does everyone get up to speed? Karen Kreeger reports.

When Daniel Hoepfner began his PhD project on apoptosis at Cold Spring Harbor Laboratory in New York, little did he know it would end up with him doing real-time imaging in Germany. The switch occurred when his research took a detour into developmental biology and his lab didn't have the relevant equipment or expertise. Fortunately, his adviser, Michael Hengartner, arranged for Hoepfner to be trained by Ralf Schnabel at the Technical University of Brunswick. What Hoepfner ended up with was much more than he anticipated — a paper in *Nature* (D. J. Hoepfner, M. O. Hengartner and R. Schnabel *Nature* **412**, 202–206; 2001) and skills that he is now putting to good use in his postdoctoral research on brain stem cells at the National Institute of Neurological Disorders and Stroke in Bethesda, Maryland.

Hoepfner's PhD experience is not unusual. Many postdocs and students find themselves scrambling to acquire fresh skills when the need arises unexpectedly — something that is happening more frequently as research becomes ever more interdisciplinary, says Michael Gottesman, deputy director for intramural research at the US National Institutes of Health (NIH). "We all realize that most scientists in the future will be part of multidisciplinary research teams," he says.

This shift is causing a change in the way that scientists need to train. Although they must still be an expert in their speciality, they must also become conversant in techniques that once seemed beyond their domain. They also need to recognize where their knowledge ends and when they should seek help from

others who have expertise in a particular field.

As co-chair of the NIH fellows committee, Hoepfner is now in a position to help other fellows navigate their way through their training needs. "As with career development, the creative nature of biomedical research often leads us to unexpected places for which we are rarely prepared," Hoepfner says.

But there are several ways to deal with one's unknown and unexpected training needs. Young scientists can seek out PhD programmes with formal training components that level the playing field for each new class. They can travel — both to other labs and to short courses. And they can engage in collaborations where they pick up new skills in an ad hoc fashion.

BACKGROUND BLENDING

For the European Molecular Biology Laboratory (EMBL) in Heidelberg, providing a mechanism to ensure that all of its PhD students can make use of its resources is essential, says Iain Mattaj, the lab's scientific director. This is because EMBL attracts people from various countries, who often have had very different kinds of training. Some have a strong theoretical background but little or no lab research experience, whereas others have spent more than a year at the bench after finishing their course work, Mattaj says.

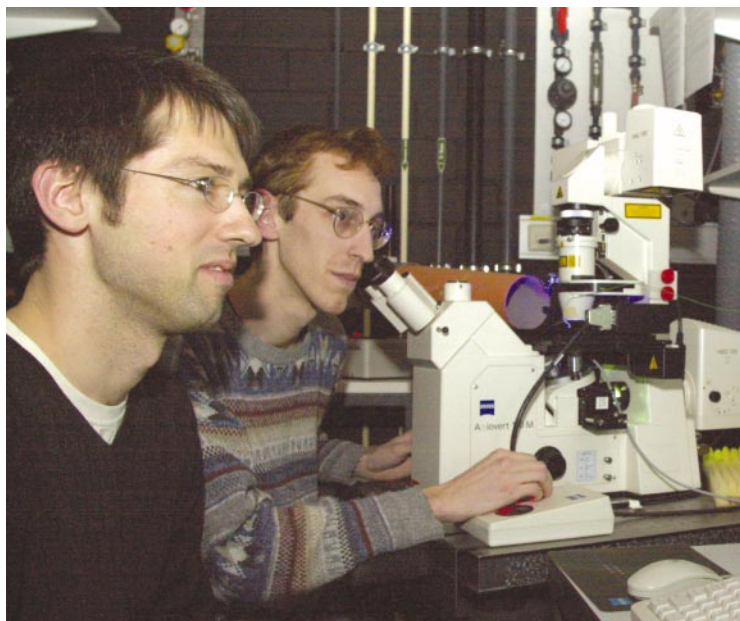
At EMBL all incoming students take an intensive introductory course on the institute's primary labs, which sees them spend time in each of the 50–60 research labs, and blends people from chemistry, physics, mathematics and biology backgrounds together. "The amazing thing is how quickly all of them get up to speed," says Mattaj. "It is difficult to tell after six to nine months who did lots of lab work and who had done none; who had very advanced molecular-biology coursework versus who had more general biology."

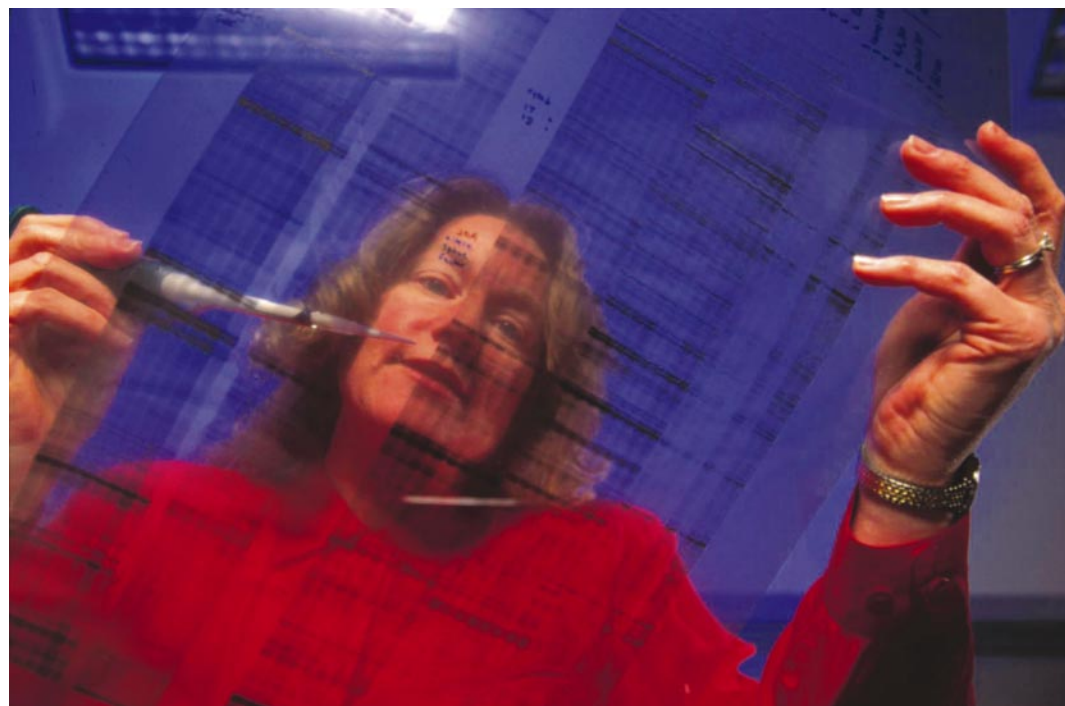
Joël Beaudouin, a PhD student in the gene-expression and cell-biology/biophysics programmes at EMBL, is an example of this effect. At the Industrial Physics and Chemistry Higher Education Institution in Paris, he studied mostly physics, with some chemistry and biology. "The culture difference with biologists was quite big at the beginning," he says. But he found that EMBL's introductory courses helped to ease the transition.

JUST VISITING

When the density of diverse labs that is found at EMBL or the NIH isn't available, many short but focused visits to other labs during graduate school work just as well. Kirsten-Peter Rabitsch, a PhD student at the Research Institute of Molecular Pathology (IMP) in Vienna, spent five years at the

Blending in: Joël Beaudouin (right) found it easy to switch from chemistry and physics to biology.





The US National Institutes of Health offers a range of courses aimed at expanding students' skills in areas such as genetic analysis.

University of Vienna studying biology and microbiology. In his fourth year he spent nine months at the University of Manchester, UK, on the European Union's Erasmus higher-education exchange programme. While there he attended lectures and practical courses in molecular biology, and worked in two labs for practical experience.

Sometimes getting up to speed means adding to your frequent-flyer miles. Despite Rabitsch's cross-training and lab experience as an undergraduate, he needed to travel back to Britain to pick up another skill. In spring 2002, while working on yeast genetics, he visited Tomoyuki Tanaka's lab at the University of Dundee for three weeks to learn a new technique — chromatin immunoprecipitation — to analyse nuclear protein complexes and their interactions with the DNA template. "Although this had been done in our lab before by several people, I couldn't get it to work," says Rabitsch. "But Kozo Tanaka, a postdoc in Tomo's lab, showed me how to do it and I got it right."

And what if you decide to switch research tactics mid-stream? When Rabitsch looked into moving on to different species of yeast as his model organism he visited Jean-Paul Javerzat's lab in Bordeaux, France, to learn about the nuances of working with this species. His travels were funded through personal money and the Erasmus programme, and his PhD adviser.

CHANGING DIRECTIONS

Many graduate students do postdoctoral research in a field that is different from their dissertation research, sometimes to pick up new skills and techniques. "If you are considering a radical change in field, you should seek out members of prospective labs who have also changed fields and ask detailed questions about their training experience," advises Hoeppner. "For example, are new postdocs paired with more senior research staff to learn the ropes?"

The next question to ask is whether the lab supports formal training through courses. Places such as the Marine Biological Laboratory in Woods Hole, Massachusetts or Cold Spring Harbor Laboratory have traditionally offered strong summer programmes. And, Hoeppner adds, it is worth asking whether the principal investigator encourages collaborations with other groups.

Unfortunately, these famous courses are often oversubscribed and have long waiting lists. Another option is perusing classes offered by large research campuses such as the NIH. For example, the Foundation for Advanced Education in the Sciences offers courses for hands-on techniques such as DNA amplification and tissue culturing, whereas the Center for Information Technology offers introductory courses on statistics, software and programming languages.

All of these avenues are great for boning up on a new area, but they can't always compare with personal collaborations. Applying the basics of an introductory statistics course to compare experimental results is one thing, "but if you plan on doing microarray analysis, you need to have a statistician on your team to give you confidence that the powerful tools in your software are being applied appropriately," says Hoeppner. Even with multidisciplinary training, you need to know your limitations and when to seek out specialist expertise, he adds. In the future, says Hoeppner, "an important skill will be recognizing the point at which you should look for a collaborator and not invest the time to develop a new system or expertise yourself".

Although there isn't any set formula for expanding your skills, a sound approach may be to be aware of your options — whether they be formal catch-up programmes, lab visits, short courses or collaborations. And, perhaps most importantly, you should also know your limitations.

Karen Kreeger is freelance science writer based in Philadelphia.

Web links

Foundation for Advanced Education in the Sciences

♦ www.faes.org

Center for Information Technology

♦ www.cit.nih.gov